National IQs: measurement and defense

- Leonardo Parra, independent researcher, sebastianxjensen@gmail.com
- Emil OW Kirkegaard, Ulster Institute for Social Research

Abstract

Using various sources of performance on cognitive tests, we constructed a set of national IQs for 197 nations, the latter using no geographic imputations. East Asian countries scored at an average of 100, Europeans at 95, Arabs at 85, Latin Americans at 80, South Asians at 75, and Sub-Saharan Africans at 70. This very low IQ of Sub-Saharan Africa in contrast to the relatively high IQ of African Americans (80-90) is incompatible with a view that national differences in intelligence are either completely environmental or genetic. Combining the various datasets reduced the estimated standard error of national IQs from 5.41 to 2.58, and a strong correlation between IQ and GDP per capita was observed (r = .82).

Based on the prior that Flynn Effect gains do not pass measurement invariance, IQ scores should exhibit some non-negligible bias between countries. Empirical assessments of measurement invariance across nations finds that measurement invariance violations are uncommon, and are more prevalent in verbal than nonverbal tests. In most countries, national IQs show high levels of reliability and validity and we encourage their use.

1. Introduction

The earliest person to compare test scores between different nations was Barbara Lerner (1983), who compared the performance of Western Europe, the United States, and Japan in test performance and hypothesized that it was related to economic development. Richard Lynn (1978; 2002) later collected IQ test scores from various countries, and found that national IQs and GDP per capita correlated at .82, though this dataset and other revisions of it have been extensively criticized in the literature. Some economists have made indexes of human capital based on child mortality, test scores, and educational attainment (Angrist et al., 2021), but it could be argued that child mortality and education are a function of both human capital and socioeconomic development, making it an improper measurement. These efforts aside, the study of national differences in intelligence has largely been a Lynn-only project.

The study of these observed national differences in intelligence by Richard Lynn has attracted quite a bit of controversy. Although there are a very large number of criticisms of this data, some valid and some not, the focal criticisms are that the Sub-Saharan African IQs are too low (Sear, 2022; Ebbesen, 2020), the use of imputations (Hunt & Sternberg, 2006), and that the selection procedure for the studies is compromised (Kamin, 2006; Sear, 2022). Providing substantial rebuttals to these criticisms is far beyond the scope of the introduction; the methodology section will contain a defense of Lynn's work and the study of measuring differences in intelligence between nations.

Underlying this debate is what exactly causes of racial differences in observed IQ scores. As races are not equally distributed across the globe, national differences in intelligence will be interpreted by some as being racial in origin, whether that be on a spiritual, cultural, environmental, or genetic level. Unsurprisingly, most of the researchers who are critical of Lynn's national IQ dataset advocate that race differences in intelligence are environmental in origin (Sear, 2022; Ebbesen, 2020; Hunt & Sternberg, 2006), and those that uncritically used the data either do not comment on the controversy (Clark et al., 2020; Rindermann, 2018) or support the theory (Templer & Arikawa, 2006). Even beyond that, there is also a moral debate on whether racial differences in intelligence should even be studied (Cofnas, 2019).

Given the controversy regarding whether there are large national differences in intelligence, we decided to create the highest quality possible measurement of national intelligence and wrote a defense of the use of national IQs, mostly in the methods section.

Sear has criticized the use of national IQs (2022), primarily the Lynn and Becker datasets for several reasons. Among these criticisms is the use of children to estimate the average IQs of nations, as IQ scores depend on age. However, the scores on these tests are standardized by age, which makes this concern irrelevant. This can be a concern if the magnitude of group differences varies by age, but the best evidence available suggests that is usually not the case, at least not between American Blacks and Whites (Rushton & Jensen, 2005). The same is true for Asians and Whites, where Asians score above Whites as children (Rushton, 1997; Weiss et al., 2019) and adults (Weiss et al., 2010). There are exceptions, such as the Arab ~ European IQ difference, where the difference increases with age (Bakhiet et al., 2018).

Sear (2022) also questions whether the figures that are estimated for the African countries are believable, as many of them fall in the 65 to 75 range, which is close to the conventional cutoff for intellectual disability (70). This ignores that not all causes and types of mental disability are the same (Jensen, 1970; Reichenberg et al., 2015): some of them are mild and typically caused by additive genetic variance, these intellectually disabled people generally can live normal lives (Boat & Wu, 2015); others are caused by severe mutations or deletions, which cause deficits in other areas of biological functioning. Arthur Jensen was initially drawn to IQ research because he noticed that Black and White children in the classes for the mentally disabled behaved quite differently in the playground, the Black children behaving normally, but the White being socially dysfunctional. The explanation for this pattern was that a large fraction of the White children suffered from major genetic disorders such as Down's syndrome, or perinatal environmental damage (syndromic disability), while the Black children were merely on the left side of their normal distribution, thus had mostly ordinary causes (familial disability). Since the syndromic causes of mental disability usually cause other deficits beyond low intelligence, this explains the large difference in the social skills of the two groups of children.

A more intuitive comparison would be differences in height between African Pygmies and those from the Dinaric Alps. On average, Pygmy men are about 153 cm tall (Travaglino et al., 2011), and Dinaric men are about 186cm tall (Pineau et al., 2005); a difference of roughly five entire standard deviations relative to the standard deviation of Dinaric male height (6.5 cm). The conventional cutoff for dwarfism in Western nations is 150cm; within the Pygmies, roughly half of their men would fall below this cutoff, in the Dinaric Alps, only men who suffer from a genetic disorder such as achondroplasia, metatropic dysplasia, or growth hormone deficiency could be this short. The fact that Dinarics who are under 150cm tall tend to suffer from additional complications that are not observed in Pygmies is not evidence that height measurements are biased against the latter group; merely that height differences must be understood as originating from a variety of genetic and environmental causes, which can have effects on various phenotypes.

It is doubtful that an IQ score of 70 for an African and a European means the same thing in terms of biological functioning, though these scores accurately reflect their ability to take cognitive tests, as Africans tend to score the equivalent of an IQ of 70 on scholastic tests administered by the TIMSS (Lynn & Meisenberg, 2010). Whether these test scores function as biased estimates of intelligence is debatable. Theoretically, some biases will deflate the African IQ relative to what would be expected from their true average levels of intelligence (low effort test takers, Flynn Effect related measurement variance, illiterates), and others will inflate it (use of primary/secondary school students which are less nationally representative in more uneducated countries, use of the standard deviation between groups instead of within groups, use of subtest differences instead of full scale differences).

Flynn Effect related measurement invariance is concerning, as the literature overwhelmingly converges towards Flynn effects being partially caused by test bias in favour of newer cohorts (Recueil, 2024; Wicherts et al., 2004; Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Pietschnig et al., 2013). As nations differ in the rate at which they undergo Flynn Effects (Pietschnig & Voracek, 2015; Rindermann & Becker, 2018), this may cause the test scores to be biased in favour of certain countries. Some of the Flynn Effect gains are still plausibly real: brain sizes increased by about 0.7 SD (DeCarli et al., 2024) between the 1930s and 70s, if this effect occurred between 1900 and 1970, then the expected increase in brain size would be 1.2 SD. Given that brain size and IQ correlate at roughly .28 (Cox et al., 2019), and this correlation is causal from brain size to intelligence (Lee et al., 2019), intelligence would have been expected to increase by 5 points due to this; assuming it is absolute and not relative brain size that is linked to IQ.

There have been some studies on whether international scholastic tests satisfy measurement invariance. There are traditionally four steps taken to test measurement invariance: configural invariance (whether the items load on the same factors between groups), metric invariance (whether the magnitude of the factor loadings on the constructs differs between groups), scalar invariance (whether the magnitude of the intercepts of the items differs between groups), and residual invariance (whether the residual variance of the items is the same between groups) (Putnick & Bornstein, 2016). For comparing national means, scalar invariance is the most important test of measurement invariance that needs to be satisfied.

Contrary to priors, scores on cognitive tests do not exhibit large violations of measurement invariance, especially if the test involved is nonverbal. Strict measurement invariance was held within Anglo and East Asian cultural groups on the 1999 TIMSS tests, though only weak (metric, but not scalar) measurement invariance was held between the cultural groups (Wu et al., 2007), as shown in Figure 1. Their methodology is limited by the fact that measurement invariance was assessed at the factor level, as groups are likely to differ in general and specific ability -- it would be better to assess measurement invariance at the item level.

Figure 1. Results of measurement invariance testing from Wu et al. 2007.

Summary Results of MI for 21 Planned Comparisons

	AUS	NZL	CAN	USA	TWN	KOR
NZL	**Strict**					
CAN	**Strict**	**Strict**				
USA	**Weak**	**Strict**	**Strict**			
TWN	Weak	Weak	Weak	Weak		
KOR	Weak	Weak	Weak	Weak	**Strict**	
JPN	Weak	Configure	Weak	Configure	**Weak**	**Strict**

Note. Results for within-culture comparisons were highlighted in bold.

The vast majority of the items on the 2015 PISA math and science tests passed measurement invariance (Odell et al., 2021), in both the factor loadings and intercepts, suggesting test bias was not an issue in administration. Another study of international test bias of the PISA item data on the reading subtest found that scalar invariance was violated in most nations, with the magnitude of invariance ranging from 0.041 in Canada to 0.93 in Kyrgyzstan (in cohen's d) (Asil & Brown, 2015). The presence of biased items, however, does not imply that the means are biased between groups, as the direction of the effects tends to vary at the item level (Cardoza, 2006; Kirkegaard, 2021).

The most exhaustive and recent assessment of measurement invariance between nations is an assessment that is available in the PISA 2022 technical report. They concluded that measurement invariance is a major issue for the creative thinking test, somewhat of an issue for the science and reading tests, and a minor issue for the mathematics test. Figures 1 and 2 show the distribution of variant (orange/red/light green) and invariant (dark green) items by country and test.

Figure 1. Results of the measurement invariance testing at the item level for the science and creative thinking by country (taken from PISA, 2022).

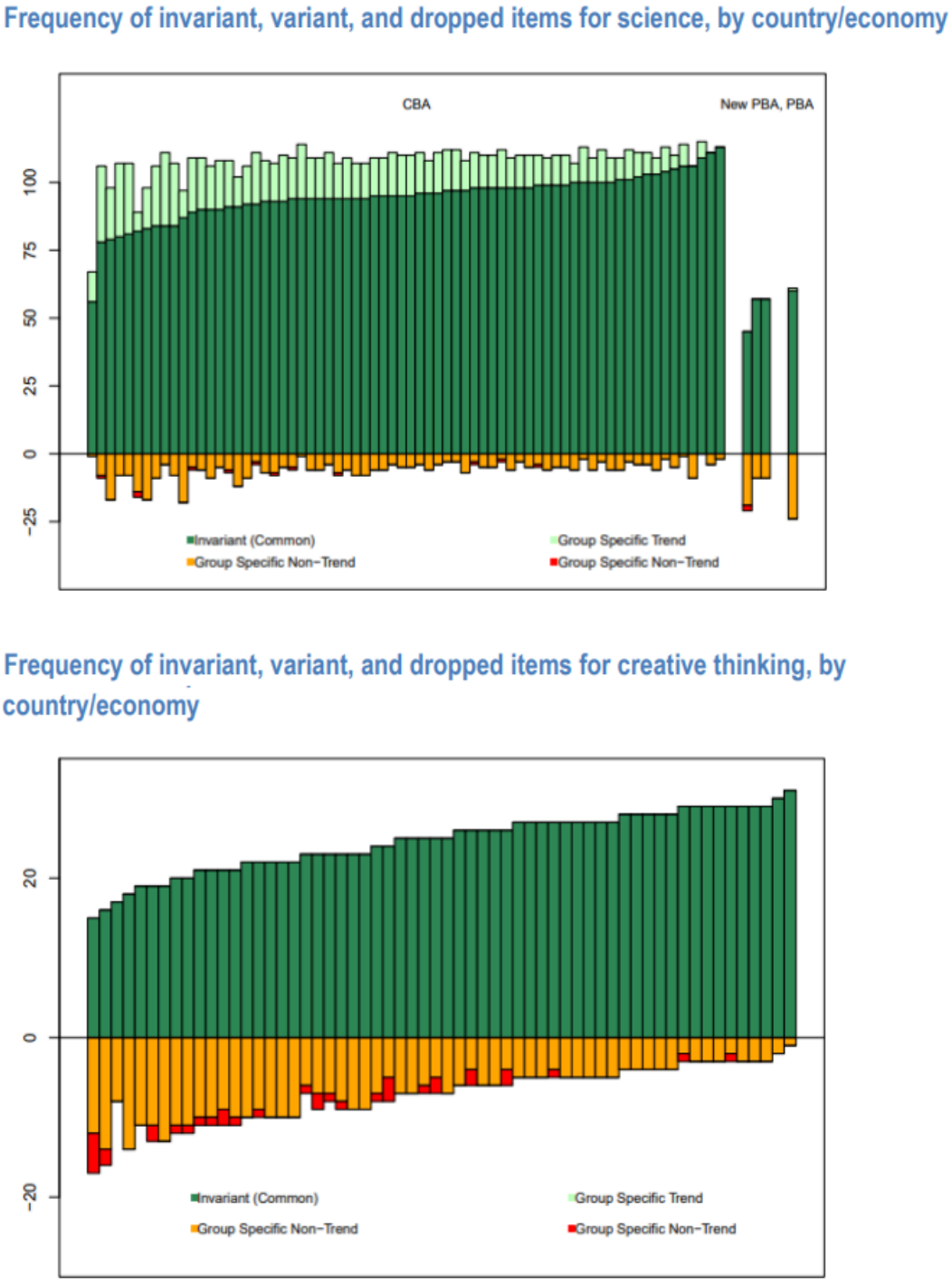
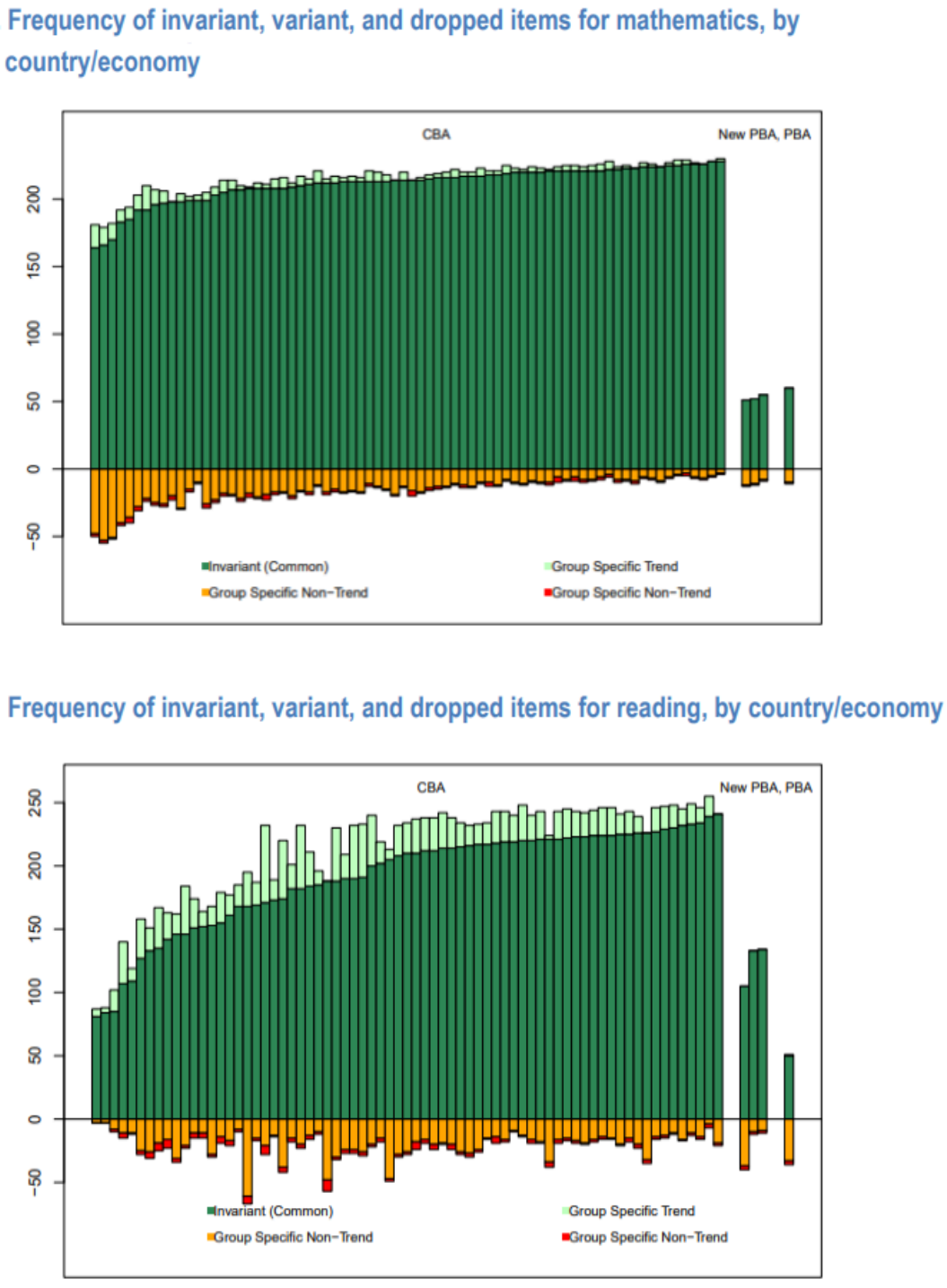


Figure 2. Results of the measurement invariance testing at the item level for the mathematics and reading test by country (taken from PISA, 2022).



In practice, the differences between countries on PISA scores are highly correlated and of roughly equal magnitude, as shown in Table 1.

Table 1. Average score on the PISA exam by country and subtest in 2022 (OECD 2022).

Country	Maths	Country	Science	Country	Reading
Singapore	575	Singapore	561	Singapore	543
Macau	552	Japan	547	Ireland	516
Chinese Taipei	547	Macau	543	Japan	516
Hong Kong	540	Chinese Taipei	537	South Korea	515
Japan	536	South Korea	528	Chinese Taipei	515
South Korea	527	Estonia	526	Estonia	511
Estonia	510	Hong Kong	520	Macau	510
Switzerland	508	Canada	515	Canada	507
Canada	497	Finland	511	United States	504
Netherlands	493	Australia	507	New Zealand	501
Ireland	492	Ireland	504	Hong Kong	500
Belgium	489	New Zealand	504	Australia	498
Denmark	489	Switzerland	503	United Kingdom	494
United Kingdom	489	Slovenia	500	Finland	490
Poland	489	United Kingdom	500	Denmark	489
Australia	487	United States	499	Poland	489
Austria	487	Poland	499	Czech Republic	489
Czech Republic	487	Czech Republic	498	Sweden	487
Slovenia	485	Denmark	494	Switzerland	483
Finland	484	Latvia	494	Italy	482
Latvia	483	Sweden	494	Germany	480
Sweden	482	Germany	492	Austria	480
New Zealand	479	Austria	491	Belgium	479
Germany	475	Belgium	491	Norway	477
Lithuania	475	Netherlands	488	Portugal	477
France	474	France	487	Croatia	475
Spain	473	Hungary	486	Latvia	475
Hungary	473	Spain	485	Spain	474
Portugal	472	Lithuania	484	France	474
Italy	471	Portugal	484	Israel	474
Vietnam	469	Croatia	483	Hungary	473
Norway	468	Norway	478	Lithuania	472
Malta	466	Italy	477	Slovenia	469
United States	465	Turkey	476	Vietnam	462
Slovakia	464	Vietnam	472	Netherlands	459

Croatia	463	Malta	466	Turkey	456
Iceland	459	Israel	465	Chile	448
Israel	458	Slovakia	462	Slovakia	447
Turkey	453	Ukraine	450	Malta	445
Brunei	442	Iceland	447	Serbia	440
Ukraine	441	Serbia	447	Greece	438
Serbia	440	Brunei	446	Iceland	436
UAE	431	Chile	444	Uruguay	430
Greece	430	Greece	441	Brunei	429
Romania	428	Uruguay	435	Romania	428
Kazakhstan	425	UAE	432	Ukraine	428
Mongolia	425	Qatar	432	Qatar	419
Cyprus	418	Romania	428	UAE	417
Bulgaria	417	Kazakhstan	423	Costa Rica	415
Moldova	417	Bulgaria	421	Mexico	415
Qatar	414	Moldova	417	Moldova	411
Chile	412	Malaysia	416	Brazil	410
Uruguay	409	Mongolia	412	Jamaica	410
Malaysia	409	Cyprus	411	Colombia	409
Montenegro	406	Colombia	411	Peru	408
Azerbaijan	397	Costa Rica	411	Montenegro	405
Mexico	395	Mexico	410	Bulgaria	404
Thailand	394	Thailand	409	Argentina	401
Peru	391	Peru	408	Panama	392
Georgia	390	Argentina	406	Malaysia	388
North Macedonia	389	Brazil	403	Kazakhstan	386
Saudi Arabia	389	Jamaica	403	Saudi Arabia	383
Costa Rica	385	Montenegro	403	Cyprus	381
Colombia	383	Saudi Arabia	390	Thailand	379
Brazil	379	Panama	388	Mongolia	378
Argentina	378	Georgia	384	Georgia	374
Jamaica	377	Indonesia	383	Guatemala	374
Albania	368	Azerbaijan	380	Paraguay	373
Indonesia	366	North Macedonia	380	Azerbaijan	365
Palestine	366	Albania	376	El Salvador	365
Morocco	365	Jordan	375	Indonesia	359
Uzbekistan	364	El Salvador	374	North Macedonia	359
Jordan	361	Guatemala	373	Albania	358
Panama	357	Palestine	369	Dominican Republic	351

Kosovo	355	Paraguay	368	Palestine	349
Philippines	355	Morocco	365	Philippines	347
Guatemala	344	Dominican Republic	360	Jordan	342
El Salvador	343	Kosovo	357	Kosovo	342
Dominican Republic	339	Philippines	356	Morocco	339
Paraguay	338	Uzbekistan	355	Uzbekistan	336
Cambodia	336	Cambodia	347	Cambodia	329

It's worth mentioning that most researchers, including Becker and Rindermann, used scholastic estimates of ability derived from international tests to estimate the intelligence of nations. These data sources are less biased than the estimates that are based on convenience samples of subjects that take IQ tests: they tend to test about a thousand students per country, the samples are roughly representative of the student body of the country, and the same test is administered to all countries at roughly the same time. The differences between the countries also cannot be attributed to collection bias. Within individuals, scores on IQ tests and scholastic ability tests correlate positively (Saß et al., 2017; Flores-Mendoza et al., 2018) and this is true between nations as well.

Some researchers have argued that the samples of Africans who took the Raven's test collected by Lynn have low levels of convergent validity and were taken from unrepresentative samples (Wicherts et al., 2010). The low scores of Africans (70) on these tests cannot be blamed on selective sampling or reporting, as the average African IQ converges to an average of roughly 70 regardless of the source (Warne, 2022), including sources that rely solely on results from scholastic assessments.

Independent of estimates based on measured IQ, the expected African *g* can be estimated based on several parameters, including the average IQ of Blacks in America, the percentage of the difference between American Blacks and Whites that is due to additive genetics, the percentage of admixture in American Blacks that is European (20%), and the extent to which the environment of Sub-Saharan Africa depresses *g* in contrast to that of America. For example, if the between-group heritability of IQ between African Americans and White Americans is 100%, and the difference in *g* between them is 18 IQ points, and the environment of Africa depresses *g* scores by 10 points, then the expected Sub-Saharan African *g* in is 67.5 (67.5 = (82-.2*100)/.8 - 10).

There is fairly robust evidence, from military-based randomization studies (Carlsson et al., 2012) and latent modeling (Karwowski & Milerski, 2021; Lasker & Kirkegaard, 2022; Ritchie et al., 2015) that education improves IQ scores, though this improvement does not translate to greater general intelligence (e.g. increases in accumulated knowledge, but not reaction time). If this conclusion is accepted, then it must be the case that differences in IQ between nations that are due to differences in educational attainment must lead to bias in favour of the more educated countries. Besides this, there is quantitative evidence summarized by Warne (2023) which indicates that unschooled populations in Central Asia do not reason about problems on IQ tests

the same way Westerners do: when asked which of a set of four objects do not fit together (e.g. an axe, saw, hammer, and log), they will typically choose one of the tools, as not much can be done without three tools and no object to operate with (Lurija, 1978).

Sear (2022) also noted that there was no formal search strategy or exclusion strategy carried out by Becker and Lynn. A fair criticism, but keep in mind that search strategies are easy to falsify and that flexibility is necessary to estimate national intelligence. In some cases, unweighted means are more accurate than sample size weighted means when the sample sizes of the studies are large, when the sample sizes are small, it would be better to weigh by the sample size; For countries that have a large amount of data (e.g. South Africa), adding psychiatric, foreign, or rural samples to the dataset would be unnecessary. In other countries that have no data available, low quality samples would be better than none. In most nations, the scholastic data is of higher quality than the psychometric data, but if the psychometric data is of high quality, then it may be wise to weigh it more highly for that specific nation.

2. Data

National IQs were sourced from various datasets, sometimes different methods used to aggregate averages from the same source. A quick overview of these sources and the number of countries they estimate the average IQ of is available in Table 2 -- most of them are different versions of Becker and Lynn's datasets or studies that assess differences between nations in scholastic ability. In a few cases, new studies were integrated into the calculation process if a country displayed unusual levels of heterogeneity in sample averages.

Table 2. Sources of variables of national differences.

Variable	Number of Countries	Time range	Source
National IQs (unweighted, psychometric)	130	1945-2017	(Becker, 2023)
National IQs (sample weighted, psychometric)	129	1945-2017	(Becker, 2023)
National IQs (quality weighted, psychometric)	130	1945-2017	(Becker, 2023)
National IQs (scholastic)	102	1945-2017	(Becker, 2023)
National IQs (composite)	148	1945-2017	(Becker, 2023)
National IQs (composite)	81	varying	(Lynn & Vanhanen, 2002)
National IQs (composite)	133	varying	(Lynn & Vanhanen, 2012)
National IQs (composite)	170	varying	(Rindermann, 2018)
Recent Test Scores (PISA, TIMSS, PIRLS)	39-81	2019-2022	(Recueil, 2023) and (Wikipedia, 2024)

Test Scores (Basic Skills Dataset -- BSD)	126	varying	(Gust et al., 2022)
Test Scores (World Bank Test Scores -- WBTS)	174	varying	(Patrinos & Angrist, 2018)
Average IQs of different countries	7	varying	(Shinwari et al., 2022), (Lynn, 2006), (Kamin, 2006), (Iliescu et al., 2016), (De La Cruz, 2022)

3. Method

3.1. National IQ standard errors

Sear's focal criticism of the national IQ datasets, particularly Lynn's and Becker's, was that the quality of data was not equally distributed across regions. This is an inevitability, given less developed countries have lower data quality, thus the criticism is not specific to intelligence measurements (World Economics, 2023). Many countries in Becker's dataset were estimated using small samples -- true, but a small sample is still better than none, and even a sample of 20 can provide a reasonably precise estimate of a population mean, as the standard error will be only 3.4 IQ points. Inaccuracy in the estimation of national averages is also not only dependent on the sample size of the individual samples, but the number of them as well.

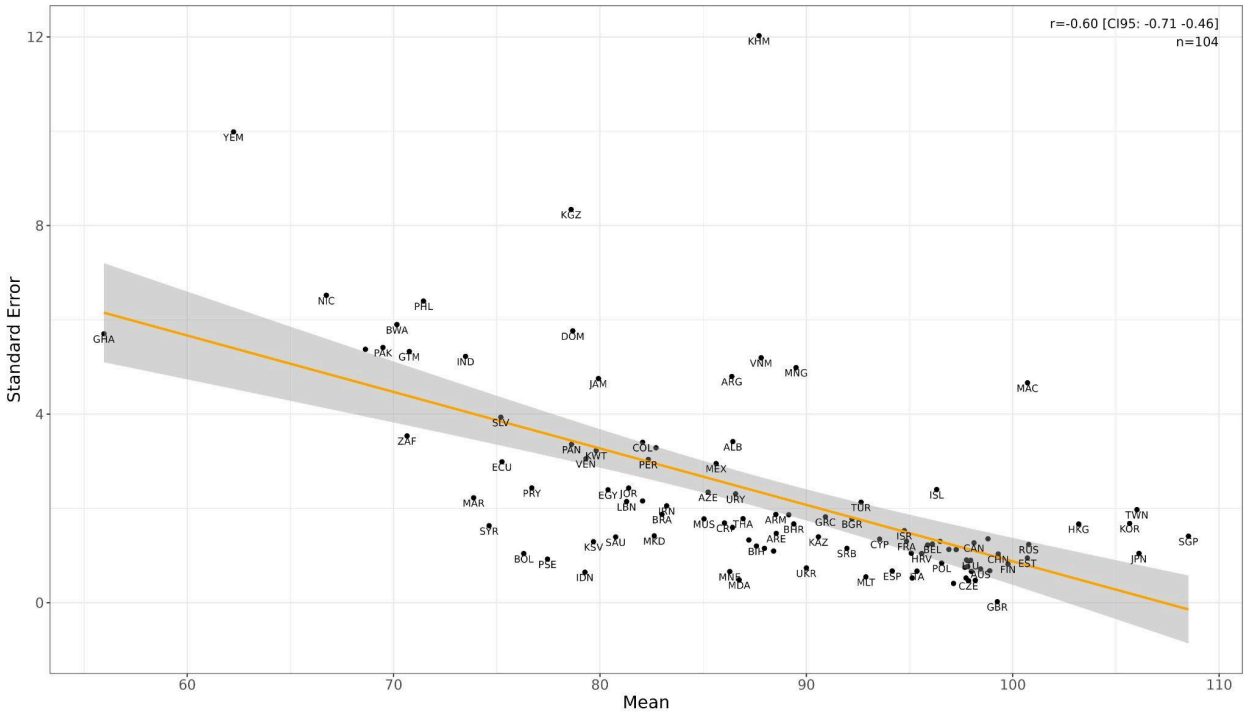
The true standard error of national IQ estimates is even higher than this, as the various proxies for national intelligence that were collected only correlated at .87 on average, implying an average standard error of 5.41 (5.41 = sqrt(1 - 0.87) x 15). This large standard error indicates that the error variance is due to heterogeneity between samples, not random sampling error. In any case, many other national datasets were based on small samples, when nothing else was available, and they were not excessively criticized for this reason (Kirkegaard & Karlin, 2020).

Warne (2022) argued in a reply to Sear that the quality of Becker's data does not vary by regional group or average level of national IQ, based on the fact that Becker's quality assessments of the data do not vary by the average IQ of the sample. This is incorrect, as high levels of sample quality in certain regions may be indicative of fraud. Empirically, Becker's quality weighted estimates of intelligence have roughly the same correlation with SDI (.81) as his unweighted estimates (.83). Based on priors, it should be the case that higher quality samples should result in more accurate estimates of intelligence; because they don't, the alternative hypothesis that the higher quality samples are more likely to be fraudulent must be considered.

The hypothesis that lower IQ nations have more imprecisely estimated means was tested by collecting estimates of national intelligence that were based on different sources of data (recent TIMSS/PIRLS/PISA assessments, Becker's psychometric estimates weighted by quality, Rindermann's estimates of scholastic ability) and estimating the means and the standard errors. The standard errors were calculated by taking the standard deviation of the sample averages and dividing them by the square root of the number of samples. Standard errors and means are correlated negatively between countries ($r = -0.60$, $p < .001$), meaning that estimates made of lower IQ countries were less accurate, as shown in Figure 3. On average, a country's estimated

IQ has a standard error of 2.33, though this figure varies substantially by country: from 0.41 in Denmark to 12 in Cambodia (the UK has 0 by default, as it is the anchor).

Figure 3. Plot of standard errors and means of national IQ estimates.



This is not due to intelligent countries having data from more samples; the negative relationship between the mean and the standard error holds after controlling for the number of samples used to estimate intelligence, as shown in Table 3.

Table 3. Regression models that predict the standard errors of the estimates.

Parameter	Model 1	Model 2	Model 3
Estimated mean IQ	-0.12 (0.016)***		-0.089 (0.018)***
Number of samples		-0.49 (0.079)***	-0.26 (0.086)**
R^2	0.36	0.28	0.41

* -> p < .05, ** -> p < .01, *** -> p < .001. Regression coefficients are unstandardized.

3.3. Estimating national intelligence averages

To compute the intelligence of nations, measured IQ and achievement test results are used. While these are not perfect measurements of intelligence, IQ scores are predictive of socially important outcomes and show low levels of bias between groups (Jensen, 1980), in contrast to personality measurements which are confounded by reference group effects (Credé et al., 2010).

Multiple sources of data were consulted, including psychometric estimates (Becker unweighted, Becker sample-weighted, Becker quality-weighted), scholastic estimates (World Bank test scores, basic skills dataset, PISA 2022 results, Becker scholastic estimates, Rindermann scholastic estimates), and composite estimates (Lynn 2012, Lynn 2002, Becker composite, Rindermann composite). If a dataset included geographic imputations, the imputations were removed.

Rindermann included estimates that were based on performance in the mathematics olympiad for North Korea, Belarus, Brunei, Cambodia, Mauritania, Tajikistan, and Turkmenistan; these were kept, though this was most relevant for Turkmenistan, which has no measured data. Samples were normed in a fashion that placed the UK at a mean of 99.26, which is roughly what the UK's average psychometric IQ is compared to British Whites. In one case where a UK sample was not available, the average of Americans was used as an anchor instead.

It was tested whether some samples were of higher quality than others, and statistical analysis suggested that this was the case (which is available in the supplement), though subjective indicators of quality (e.g. how new the data is, how much data the indicators are based on) was also taken into consideration. Concretely speaking, Lynn's and Becker's composite estimates were given lower weights due to the fact that they are based on older data and provide little incremental validity. An overall average was computed using nested means:

- Nest 1: Lynn's estimates, Becker's composite estimates, Becker's scholastic estimates, and recent TIMSS math results.
- Nest 2: average of nest 1, recent TIMSS science results, average of Becker's psychometric estimates, recent PIRLS results, World Bank test scores
- Nest 3: average of nest 2, recent PISA results, and Rindermann's scholastic estimates
- Nest 4: average of nest 3, basic skills dataset, Rindermann's IQ estimates

Another method was tested where random effects meta-analytic means were calculated for each country. Sample sizes were assigned based on the perceived quality of each dataset:

N = 10 → TIMSS math, Becker psychometric averages
N = 20 → Becker composite, TIMSS science, Lynn estimates, Becker's scholastic estimates
N = 40 → PIRLS results, PISA results, WB test scores, Rindermann SAS estimates
N = 80 → Rindermann IQ estimates and Basic Skills dataset

Samples that displayed unusual heterogeneity or extreme means in either direction were manually reviewed, where the sources were consulted and a subjective best estimate was given. Most countries that had suspiciously large amounts of variance in estimates were undeveloped countries, though there were notable exceptions like Vietnam and China. In the case of Vietnam, Becker included estimates of the IQ of rural Vietnamese who scored an IQ of 78 in his dataset; their performance on the PISA tests suggests that the true national IQ is somewhere between 95 and 100. In China, the differences in estimates between datasets is due to a debate over how the PISA samples should be weighted relative to the rest of China.

The World Bank estimated its human capital to be the IQ equivalent of 90, while the Basic Skills Dataset estimated its human capital to be the IQ equivalent of 107 -- both agreed that the PISA results were not representative, but differed in the extent to which this biased the overall average. Using the China Family Panel Study (CFPS, 2020), regional differences in cognitive ability were calculated, and it was determined that China's recent PISA results are biased because they come from more intelligent provinces like Shanghai (IQ = 107) and Beijing (IQ = 108), and that if the results were weighted relative to the whole population, they are indicative of an IQ of roughly 99. The scores from the IQ samples are also inflated by the fact that they come from educated and Eastern samples, when this bias is corrected for, the results imply an average of roughly 102 for the whole country.

In total, 42 countries had their national IQs estimated based on a manual review, and the estimates correlated at .97 with the estimates that would have been made otherwise and were 1.9 IQ points higher ($p < .001$, two-sided paired t-test) on average. In most cases, the manual revisions were unnecessary, as shown in Table 4.

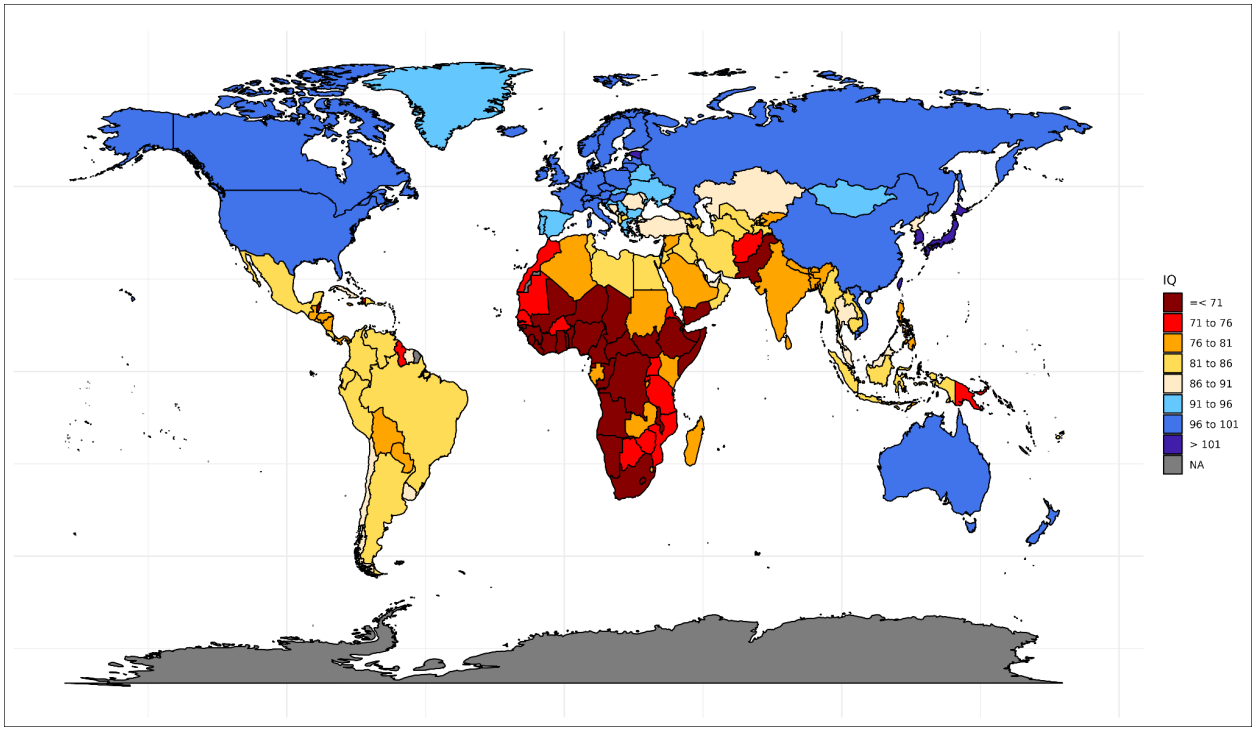
Table 4. Average IQ by country, by method.

Country	Mathematical estimate	Manual (final) estimate
Afghanistan	74.80	75.70
Cambodia	83.09	84.10
Canada	100.22	100.88
China	101.03	100.20
Cuba	90.64	87.90
Dominica	68.96	75.84
Dominican Republic	77.07	82.41
Ecuador	80.50	82.04
Egypt	79.56	81.26
El Salvador	77.14	79.87
Equatorial Guinea	61.56	69.67
Estonia	101.14	101.86
Finland	100.62	100.86
Gambia	62.83	63.70
Guatemala	75.46	78.78
Haiti	71.89	72.74
Honduras	74.57	79.30
Hong Kong SAR China	103.54	106.02
Iraq	84.62	82.27
Ireland	98.02	99.10
Jamaica	77.18	79.82
Japan	103.96	105.90
North Korea		87.90

South Korea	104.00	103.84
Kuwait	79.51	84.26
Kyrgyzstan	77.29	80.51
Laos	84.23	84.77
Macao SAR China	102.62	103.90
Marshall Islands	80.45	86.50
Mongolia	89.66	93.37
Nepal	73.01	76.98
Netherlands	99.58	100.08
Nicaragua	74.39	77.95
Pakistan	73.42	70.86
Papua New Guinea	79.37	71.77
Romania	89.14	87.34
Samoa	81.91	88.00
Singapore	106.37	108.70
Taiwan	103.34	105.23
Uzbekistan	83.88	83.95
Vietnam	93.63	98.52
Zambia	70.52	77.00

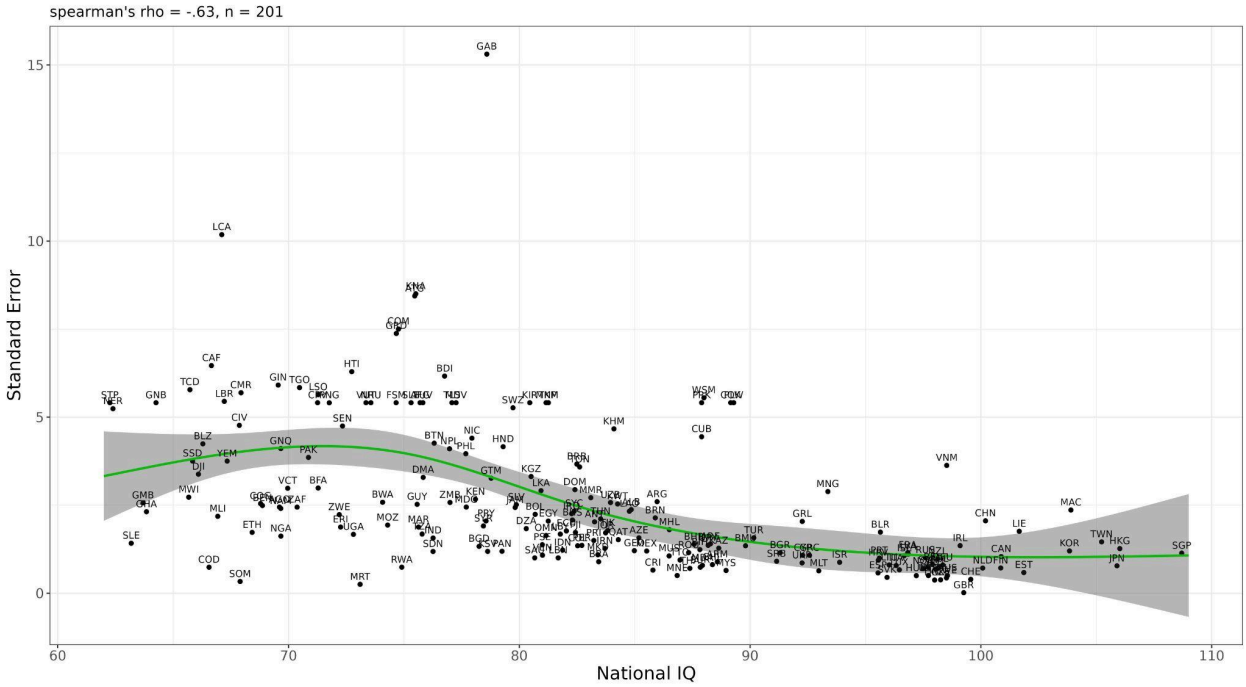
The average IQ of the world is 85.3 when weighted by population size; Figure 4 plots the means by country graphically.

Figure 4. IQ by country.



The analysis that related the standard errors and the means of national IQs was repeated for the dataset that included all national IQ datasets. We found a negative correlation between standard errors and means (spearman's rho = -.63, $p < .001$), meaning that countries with higher IQs had their estimates more precisely taken, as shown in Figure 5. This negative correlation also held for socioeconomic development, where more developed countries had lower standard errors (rho = -.65, $p < .001$).

Figure 5. Relationship between standard errors of national IQs and estimated national IQ.



4. Discussion

The national IQ estimates were shown to have non-negligible inaccuracy -- a standard error of roughly 5.41 IQ points. We have estimated that the composite measurement (SE of 2.6) has 50% less error than the average dataset that measures proxies for national intelligence. Most of the estimates made of individual countries are accurate, though a few have very high standard errors (Gabon, Cambodia, Cuba, Saint Lucia, and Haiti) or are based on dubious estimation methods (Turkmenistan was estimated using mathematical olympiad performance, North Korea was estimated using North Korean refugees and it was difficult to judge how to correct for Flynn Effects). We also found that more intelligent and developed countries tended to have more precisely estimated national IQs, even after controlling for the fact that intelligent and developed countries are more likely to be represented in these datasets.

The research on whether scholastic test scores between nations pass measurement invariance suggests that measurement invariance between countries is usually tenable, with nonverbal tests (e.g. mathematics) showing more invariance (i.e. being better comparable) than verbal (e.g. reading) ones. As these nonverbal and verbal tests have differences of roughly the same magnitude across countries, the violations of measurement invariance are not likely to be a practically significant source of bias when assessing differences in IQ between countries. Some studies have suggested that matrix reasoning does not test intelligence equally between Europeans and Sub-Saharan Africans -- the research is not definitive enough to make inferences, unfortunately.

Some groups that are similar in ancestry still differ greatly in IQ: South Koreans score 16 points higher than North Korean refugees on cognitive tests, and African Americans score 11-14 points higher than Africans. This sets a rough upper limit on how much Flynn Effects can bias estimates of intelligence between nations. The magnitude of the observed differences between nations is much larger than this, with scores ranging from 108.7 in Singapore to 62.26 in Sao Tome. Because of that, it would be rational to conclude that the disparities in test scores between countries are largely due to true differences in ability instead of test bias.

5. Appendix

Table A1. Estimated mean and standard error of IQ by country

Rank	Country	IQ	Standard Error
1	Singapore	108.70	1.14
2	Hong Kong SAR China	106.02	1.27
3	Japan	105.90	0.78
4	Taiwan	105.23	1.46
5	Macao SAR China	103.90	2.36
6	South Korea	103.84	1.20
7	Estonia	101.86	0.59
8	Liechtenstein	101.66	1.76
9	Canada	100.88	1.04
10	Finland	100.86	0.72
11	China	100.20	2.06
12	Netherlands	100.08	0.71
13	Switzerland	99.56	0.40
14	United Kingdom	99.26	0.02
15	Ireland	99.10	1.35
16	Australia	98.55	0.50
17	Vietnam	98.52	3.63
18	Sweden	98.51	0.44
19	Germany	98.35	0.80
20	Czechia	98.25	0.38
21	Poland	98.19	0.76
22	New Zealand	98.13	0.98
23	Austria	98.05	0.68
24	Denmark	98.00	0.37
25	Belgium	97.90	0.82
26	United States	97.73	0.50
27	Slovenia	97.72	0.73
28	Russia	97.59	1.01

29	Norway	97.50	0.70
30	Hungary	97.20	0.50
31	Latvia	96.85	1.11
32	France	96.83	1.15
33	Iceland	96.68	1.06
34	Luxembourg	96.47	0.67
35	Italy	96.33	0.79
36	Lithuania	96.03	0.80
37	Slovakia	95.93	0.45
38	Belarus	95.64	1.74
39	Portugal	95.60	1.00
40	Croatia	95.55	0.94
41	Spain	95.54	0.58
42	Israel	93.88	0.88
43	Mongolia	93.37	2.89
44	Malta	92.98	0.64
45	Greece	92.57	1.08
46	Cyprus	92.28	1.08
47	Greenland	92.26	2.04
48	Ukraine	92.25	0.86
49	Bulgaria	91.30	1.15
50	Serbia	91.15	0.91
51	Turkey	90.16	1.57
52	Bermuda	89.80	1.35
53	Palau	89.29	5.41
54	Cook Islands	89.16	5.41
55	Malaysia	88.96	0.65
56	Kazakhstan	88.64	1.28
57	Armenia	88.58	0.89
58	Chile	88.37	0.81
59	United Arab Emirates	88.28	1.40
60	Uruguay	88.18	1.36
61	Samoa	88.00	5.55
62	Moldova	87.93	0.79
63	Cuba	87.90	4.44
64	North Korea	87.90	5.41
65	Suriname	87.84	0.74
66	Bosnia & Herzegovina	87.82	1.24
67	Bahrain	87.57	1.39

68	Thailand	87.39	0.71
69	Romania	87.34	1.15
70	Trinidad & Tobago	86.96	0.94
71	Montenegro	86.84	0.50
72	Marshall Islands	86.50	1.81
73	Mauritius	86.49	1.06
74	Argentina	85.97	2.60
75	Brunei	85.89	2.14
76	Costa Rica	85.79	0.66
77	Mexico	85.52	1.20
78	Azerbaijan	85.18	1.57
79	Georgia	84.99	1.21
80	Albania	84.85	2.38
81	Laos	84.77	2.33
82	Qatar	84.29	1.52
83	Kuwait	84.26	2.53
84	Cambodia	84.10	4.67
85	Uzbekistan	83.95	2.58
86	Tajikistan	83.83	1.78
87	Jordan	83.74	1.71
88	Iran	83.71	1.27
89	Tunisia	83.52	2.11
90	Brazil	83.44	0.89
91	North Macedonia	83.41	1.09
92	Puerto Rico	83.23	1.50
93	Myanmar (Burma)	83.10	2.71
94	Peru	82.71	1.36
95	Tonga	82.61	3.59
96	Colombia	82.53	1.35
97	Barbados	82.49	3.67
98	Fiji	82.43	1.73
99	Dominican Republic	82.41	2.94
100	Seychelles	82.38	2.33
101	Bahamas	82.30	2.08
102	Iraq	82.27	2.26
103	Ecuador	82.04	1.77
104	Indonesia	81.88	1.22
105	Libya	81.78	1.68
106	Lebanon	81.69	1.00

107	Turkmenistan	81.26	5.41
108	Egypt	81.26	2.05
109	Northern Mariana Islands	81.16	5.41
110	Oman	81.16	1.64
111	Venezuela	81.00	1.08
112	Palestinian Territories	81.00	1.38
113	Sri Lanka	80.94	2.91
114	Bolivia	80.69	2.24
115	Saudi Arabia	80.67	1.00
116	Kyrgyzstan	80.51	3.31
117	Kiribati	80.45	5.41
118	Algeria	80.30	1.83
119	El Salvador	79.87	2.51
120	Jamaica	79.82	2.44
121	Eswatini	79.73	5.27
122	Honduras	79.30	4.16
123	Panama	79.25	1.19
124	Guatemala	78.78	3.27
125	Kosovo	78.63	1.19
126	Gabon	78.59	15.31
127	Paraguay	78.56	2.05
128	Syria	78.45	1.91
129	Bangladesh	78.26	1.33
130	Kenya	78.10	2.67
131	Nicaragua	77.95	4.40
132	Madagascar	77.70	2.45
133	Philippines	77.68	3.96
134	Maldives	77.26	5.41
135	Timor-Leste	77.08	5.41
136	Zambia	77.00	2.58
137	Nepal	76.98	4.10
138	Burundi	76.76	6.17
139	Bhutan	76.31	4.26
140	India	76.27	1.57
141	Sudan	76.26	1.19
142	Dominica	75.84	3.29
143	Tanzania	75.79	1.68
144	Afghanistan	75.70	5.41
145	Morocco	75.63	1.88

146	Guyana	75.57	2.52
147	St. Kitts & Nevis	75.52	8.50
148	Antigua & Barbuda	75.47	8.44
149	Solomon Islands	75.31	5.41
150	Rwanda	74.91	0.74
151	Comoros	74.77	7.50
152	Grenada	74.67	7.38
153	Mozambique	74.30	1.94
154	Botswana	74.08	2.58
155	Nauru	73.57	5.41
156	Vanuatu	73.36	5.41
157	Mauritania	73.10	0.25
158	Uganda	72.81	1.67
159	Haiti	72.74	6.29
160	Senegal	72.34	4.75
161	Eritrea	72.26	1.88
162	Zimbabwe	72.20	2.23
163	Papua New Guinea	71.77	5.41
164	Burkina Faso	71.29	2.99
165	Lesotho	71.29	5.64
166	Cape Verde	71.26	5.41
167	Pakistan	70.86	3.86
168	Togo	70.48	5.84
169	South Africa	70.37	2.45
170	St. Vincent & Grenadines	69.97	2.98
171	Nigeria	69.67	1.62
172	Equatorial Guinea	69.67	4.11
173	Namibia	69.67	2.41
174	Angola	69.61	2.45
175	Guinea	69.55	5.91
176	Benin	68.87	2.49
177	Congo - Brazzaville	68.79	2.54
178	Ethiopia	68.42	1.73
179	Cameroon	67.94	5.69
180	Somalia	67.90	0.33
181	Côte d’Ivoire	67.87	4.77
182	Yemen	67.34	3.75
183	Liberia	67.22	5.45
184	St. Lucia	67.11	10.18

185	Mali	66.93	2.18
186	Central African Republic	66.66	6.46
187	Congo - Kinshasa	66.56	0.74
188	Belize	66.29	4.24
189	Djibouti	66.10	3.38
190	South Sudan	65.84	3.75
191	Chad	65.73	5.78
192	Malawi	65.68	2.73
193	Guinea-Bissau	64.26	5.41
194	Ghana	63.85	2.32
195	Gambia	63.70	2.57
196	Sierra Leone	63.18	1.42
197	Niger	62.40	5.24
198	São Tomé & Príncipe	62.26	5.41

Figure A1. IQ by country, Europe only.

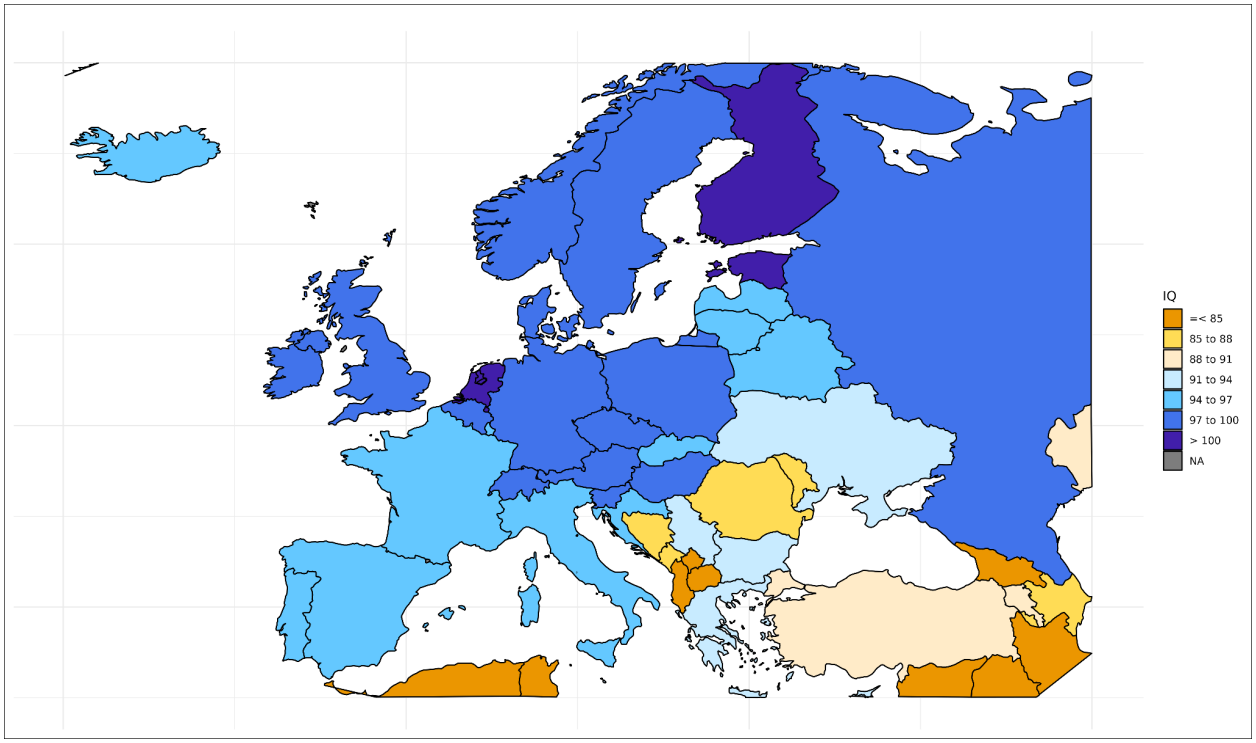


Figure A2. Relationship between national IQ (Lynn, 2002) and national IQ (estimated in 2024).

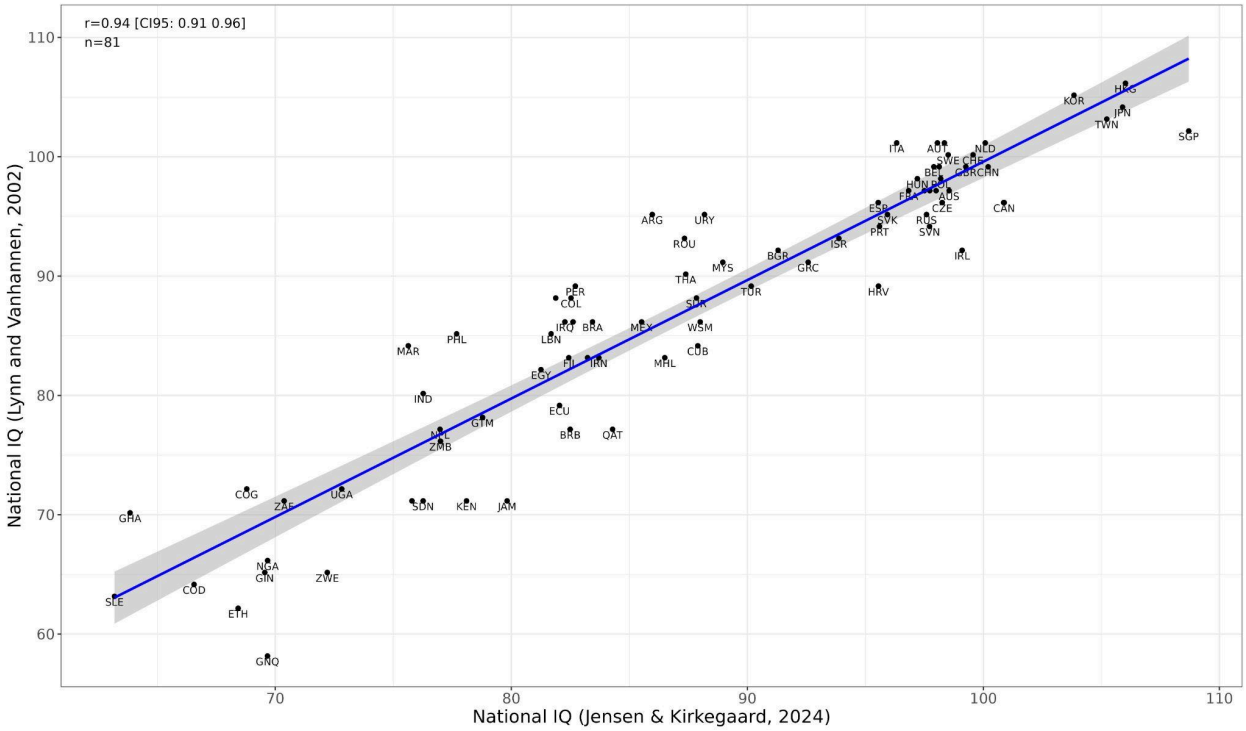


Figure A3. Relationship between national IQ (Lynn & Vanhanen, 2002) and world bank harmonized test scores (estimated in 2010-2020, converted to IQ units).

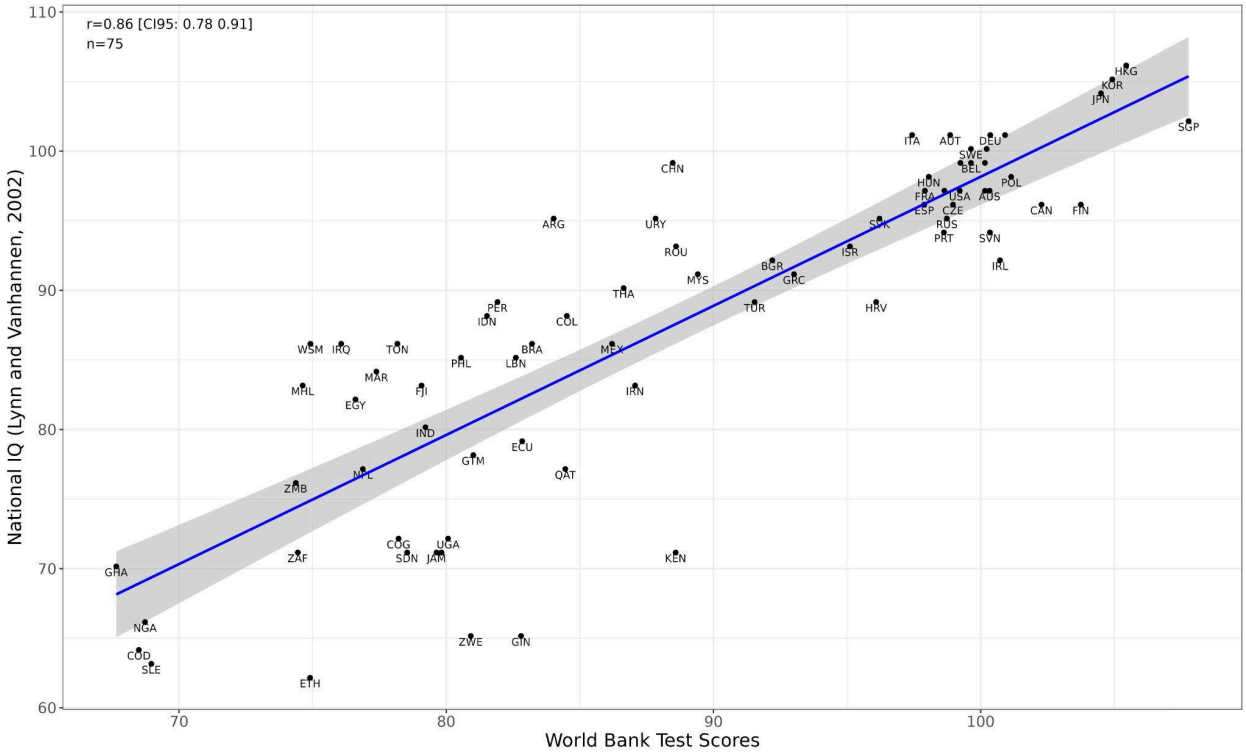


Figure A4. Relationship between average IQ and GDP per capita. Yellow line - linear fit, blue line - Locally estimated scatterplot smoothing. R^2 of the linear fit = 52%, R^2 of the nonlinear fit = 70%. GDP per capita here is an average calculated from four sources (World Bank Open Data, 2023; IMF, 2024; CIA, 2023; United Nations, 2023) and 5 years 2018-2022.

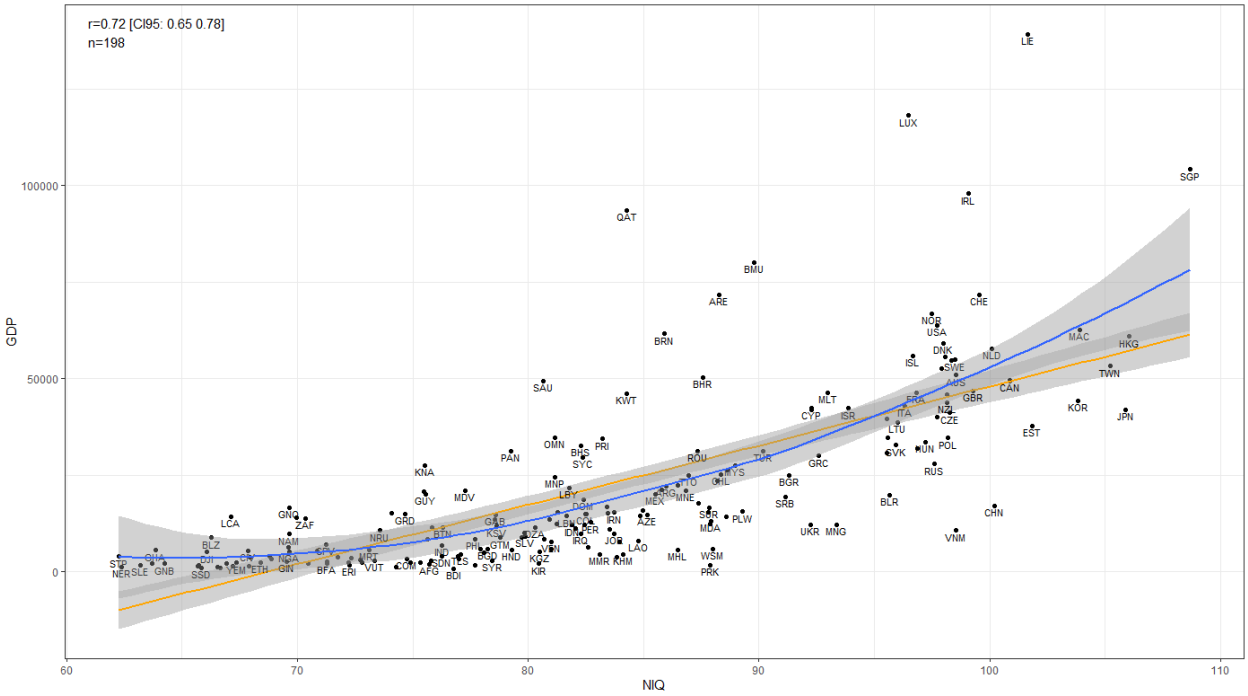


Figure A5. Relationship between predicted % who score above 125 and GDP per capita. Yellow line - linear fit, blue line - Locally estimated scatterplot smoothing. R^2 of the linear fit = 46%, R^2 of the nonlinear fit = 57%.

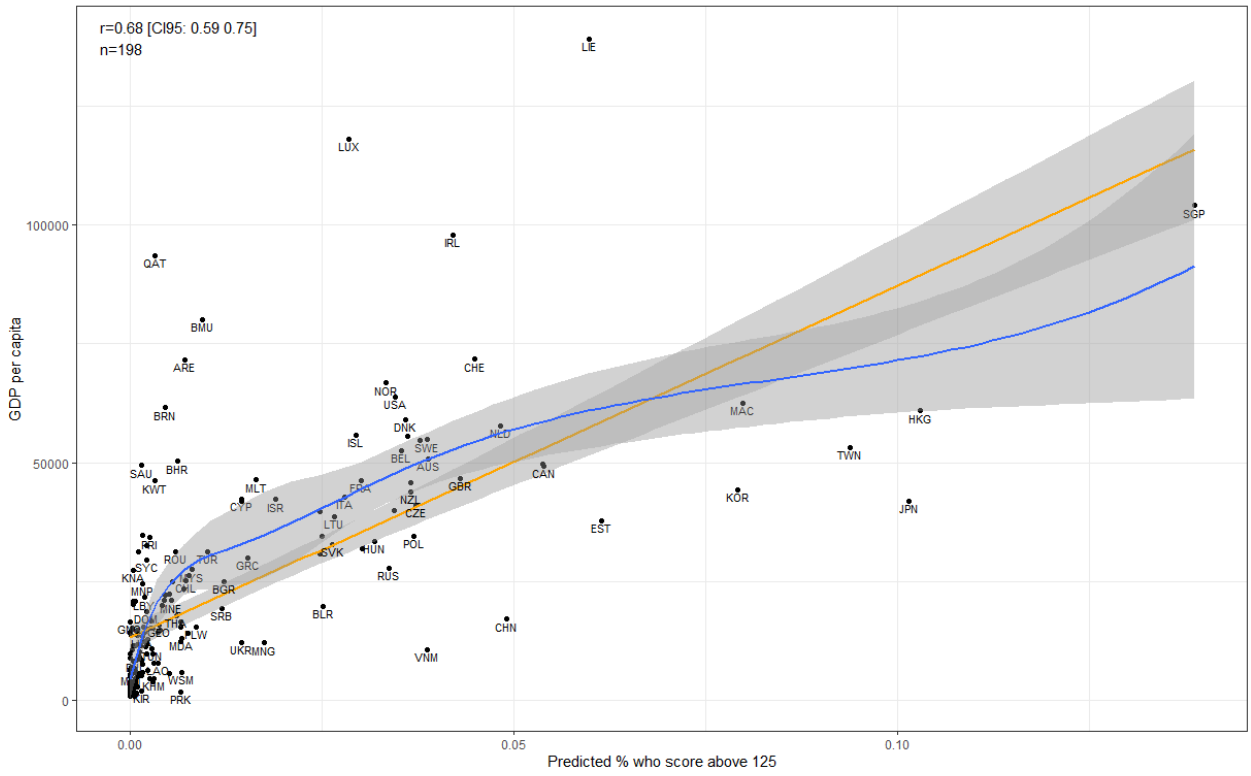


Figure A6. Hierarchical cluster analysis of the Gower distance of nations based on 47 socioeconomic development variables.

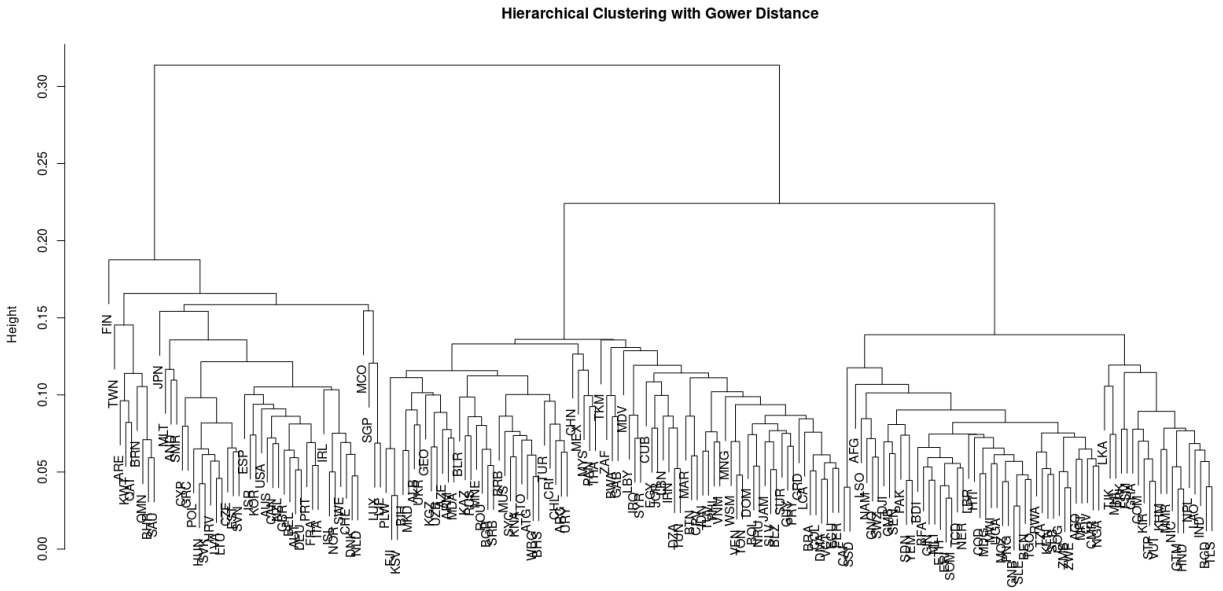


Figure A8. IQ by country (alternative colours).

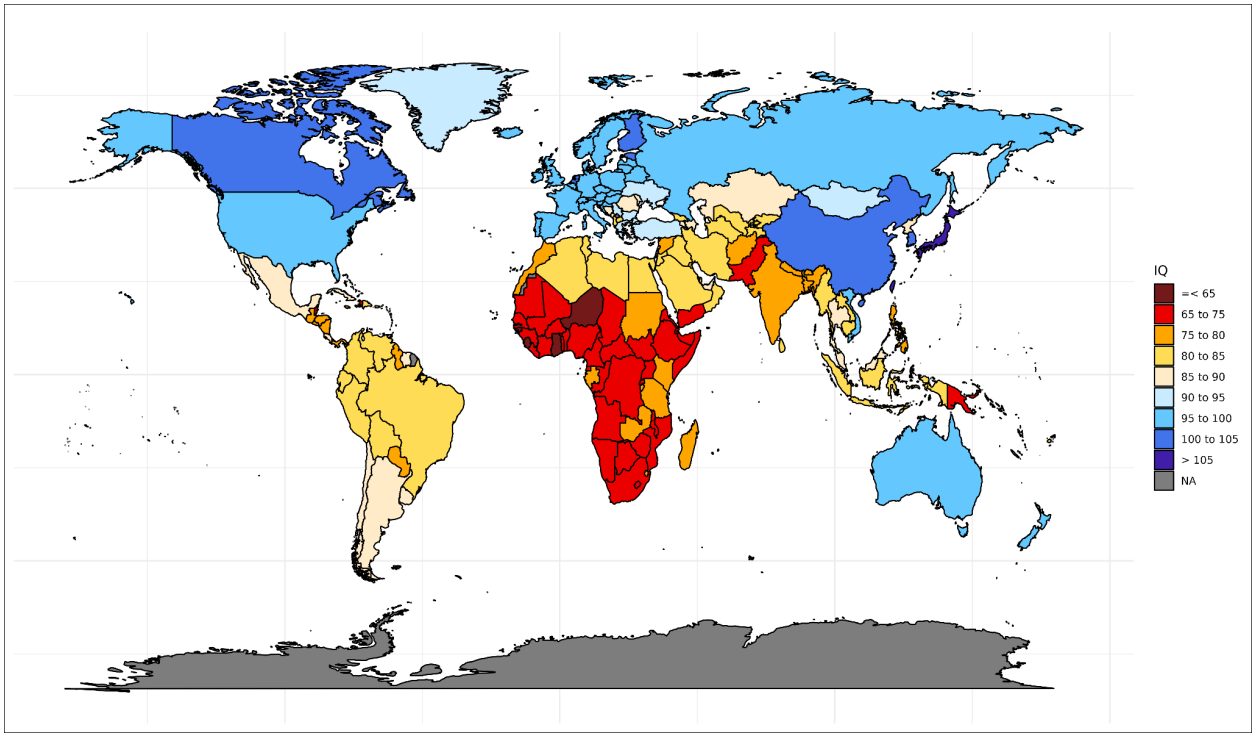


Figure A10. Relationship between log(GNI) and national IQ. The formal equation is log(GNI) = 0.0876*NIQ + 2.09. An increase in IQ of one unit corresponds to an increase in GNI per capita of 9.2%.

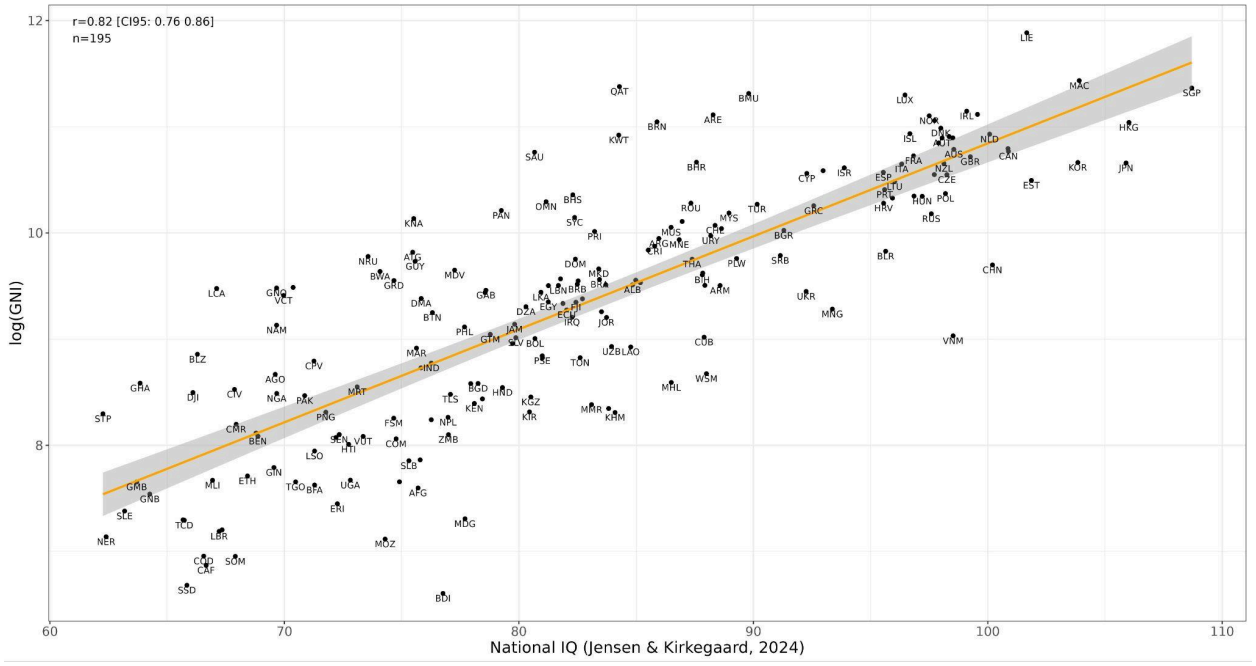
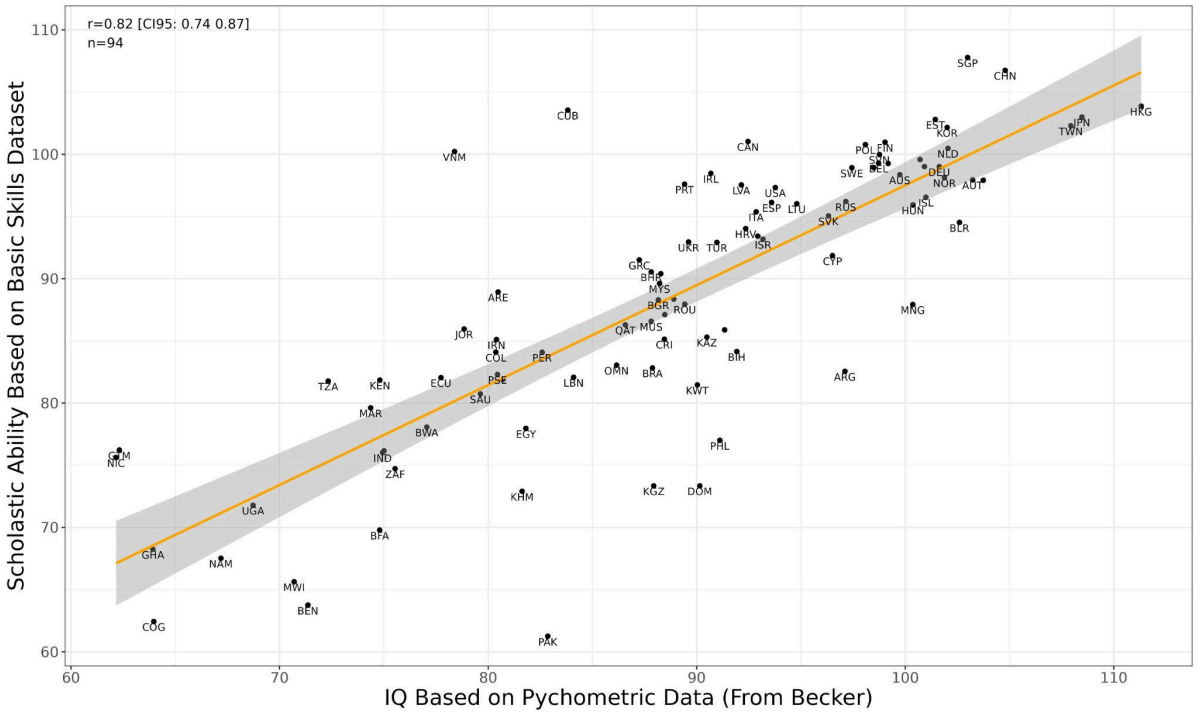


Table A2. Estimated regional IQ by dataset. BSD - basic skills dataset, WBTS - world bank test scores, RSAS - Rindermann's scholastic estimates, BSAS - Becker's scholastic estimates, BQNW - Becker's quality weighted psychometric estimates, BNW - Becker's sample size weighted estimates, BUW - Becker's unweighted estimates, SCH - average of the scholastic estimates (BSD, WBTS, RSAS, BSAS), PSY - average of the psychometric estimates (BNW, BUW, BQNW).

Region	BSD	WBTS	RSAS	BSAS	BQNW	BNW	BUW	SCH	PSY
Eastern Asia	101.76	98.89	97.51	100.63	103.37	103.27	105.81	99.70	104.15
Northern America	99.18	100.75	98.76	99.23	95.55	95.62	93.84	99.48	95.00
Western Europe	99.16	99.16	98.12	98.68	100.23	99.83	101.68	98.78	100.58
Northern Europe	98.76	99.80	97.86	98.33	96.98	96.72	97.61	98.69	97.10
Australia and New Zealand	98.68	100.25	98.26	97.71	100.07	100.03	100.33	98.72	100.14
Eastern Europe	93.76	94.95	93.26	94.98	93.24	93.18	95.22	94.24	93.88
Southern Europe	90.80	91.55	90.01	90.66	91.60	91.52	91.93	90.75	91.68
South-eastern Asia	88.11	87.42	85.76	88.61	89.10	88.98	87.24	87.47	88.44
Western Asia	86.31	85.03	79.32	79.69	83.28	83.15	84.97	82.59	83.80
Latin America / Caribbean	82.48	82.01	75.41	78.18	81.29	80.99	81.51	79.52	81.26
Central Asia	79.32	88.93	78.76	81.52	86.98	86.98	89.29	82.13	87.75
Northern Africa	79.19	78.21	75.51	72.09	78.21	78.17	78.27	76.25	78.22
Southern Asia	74.12	78.54	74.26	76.62	76.44	76.33	78.22	75.88	76.99
Sub-Saharan Africa	70.32	77.71	65.93	66.54	69.60	69.51	70.30	70.12	69.80

Figure A11. Relationship between measured IQ and scholastic ability by country. 95% confidence interval of the regression line is highlighted in grey.



The relationship between IQ based on psychometric data and scholastic estimates also holds within regions, although the relationship attenuated (r = .41, weighted by sample size), as shown in Table 5. This indicates that this correlation is not a function of regions being assigned systematically lower or higher values by the data sources, rather that nations differ in ability, and these differences are reflected in test performance.

Table 5. Correlation between Becker's unweighted estimates of IQ and the world bank test score results by region. World bank test scores were used over the basic skills dataset because the world bank dataset measured more nations. Sample size here denotes the number of countries.

Region	Correlation	Sample Size
Central Asia	0.97	4
Sub-Saharan Africa	0.70	22
Eastern Europe	0.66	8
Eastern Asia	0.60	5
Western Asia	0.58	14
Southern Europe	0.44	9
South-eastern Asia	0.39	8
Latin America / Caribbean	0.29	15
Southern Asia	0.23	6
Northern Europe	0.07	10

Northern Africa	-0.41	3
Western Europe	-0.49	6

6. References

Angrist, N., Djankov, S., Goldberg, P. K., & Patrinos, H. A. (2021). Measuring human capital using global learning data. *Nature*, *592*(7854), 403–408. https://doi.org/10.1038/s41586-021-03323-7

Asil, M., & Brown, G. T. L. (2015). Comparing OECD PISA reading in english to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, *16*(1), 71–93. https://doi.org/10.1080/15305058.2015.1064431

Bakhiet, S. F. A., Dutton, E., Ashaer, K. Y. A., Essa, Y. A. S., Blahmar, T. A. M., Hakami, S. M., & Madison, G. (2018). Understanding the Simber Effect: Why is the age-dependent increase in children's cognitive ability smaller in Arab countries than in Britain? *Personality and Individual Differences*, *122*, 38–42. https://doi.org/10.1016/j.paid.2017.10.002

Beaujean, A. A., & Osterlind, S. J. (2008). Using Item Response Theory to assess the Flynn Effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence*, *36*(5), 455–463. https://doi.org/10.1016/j.intell.2007.10.004

Beaujean, A., & Sheng, Y. (2010). Examining the Flynn Effect in the General Social Survey Vocabulary test using item response theory. *Personality and Individual Differences*, *48*(3), 294–298. https://doi.org/10.1016/j.paid.2009.10.019

Becker, D. (2023). The NIQ-dataset (V1.3.5). Chemnitz, Germany

Boat, T., & Wu, J. (2015). *Mental disorders and disabilities among low-income children*. National Academies Press. http://dx.doi.org/10.17226/21780

Cardoza, S. (2006). *Differential Item Functioning in Asian-Americans on the Stanford-Binet Standardization Fifth Edition Verbal Subtests* [Doctor of Psychology (PsyD)]. https://digitalcommons.georgefox.edu/psyd/464

Carlsson, M., Dahl, G., & Rooth, D.-O. (2012). *The effect of schooling on cognitive skills*. National Bureau

of Economic Research. http://dx.doi.org/10.3386/w18484

CFPS. (2020). China Family Panel Studies. https://www.isss.pku.edu.cn/cfps/en/data/public/index.htm

CIA. (2023). *Real GDP per capita*. https://www.cia.gov/the-world-factbook/field/real-gdp-per-capita/country-comparison/

Clark, C. J., Winegard, B. M., Beardslee, J., Baumeister, R. F., & Shariff, A. F. (2020). RETRACTED: Declines in religiosity predict increases in violent crime—but not among countries with relatively high average IQ. *Psychological Science*, *31*(2), 170–183. https://doi.org/10.1177/0956797619897915

Cofnas, N. (2019). Research on group differences in intelligence: A defense of free inquiry. *Philosophical Psychology*, *33*(1), 125–147. https://doi.org/10.1080/09515089.2019.1697803

Cox, S. R., Ritchie, S. J., Fawns-Ritchie, C., Tucker-Drob, E. M., & Deary, I. J. (2019). Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence*, *76*, 101376. https://doi.org/10.1016/j.intell.2019.101376

Credé, M., Bashshur, M., & Niehorster, S. (2010). Reference group effects in the measurement of personality and attitudes. *Journal of Personality Assessment*, *92*(5), 390–399. https://doi.org/10.1080/00223891.2010.497393

De La Cruz, J. (2022). *RELATIONSHIP AMONG DENTAL FLUOROSIS, INTELLECTUAL QUOTIENT AND ACADEMIC PERFORMANCE*. ProQuest. https://www.proquest.com/openview/71d8029de3c3adadc3d6c5afaea1c304/1?pq-origsite=gscholar&cbl=2045919

DeCarli, C., Maillard, P., Pase, M. P., Beiser, A. S., Kojis, D., Satizabal, C. L., Himali, J. J., Aparicio, H. J., Fletcher, E., & Seshadri, S. (2024). Trends in intracranial and cerebral volumes of Framingham heart study participants born 1930 to 1970. *JAMA Neurology*, *81*(5), 471. https://doi.org/10.1001/jamaneurol.2024.0469

Ebbesen, C. L. (2020). *Flawed estimates of cognitive ability in Clark et al. Psychological Science, 2020*. Center for Open Science. https://doi.org/10.31234/osf.io/tzr8c

Flores-Mendoza, C., Ardila, R., Rosas, R., Lucio, M. E., Gallegos, M., & Colareta, N. R. (2018). *Intelligence measurement and school performance in Latin America: A report of the Study of Latin*

American Intelligence Project. Springer.

Gust, S., Hanushek, E., & Woessmann, L. (2022). *Global universal basic skills: Current deficits and implications for world development*. National Bureau of Economic Research. http://dx.doi.org/10.3386/w30566

Hunt, E., & Sternberg, R. J. (2006). Sorry, wrong numbers: An analysis of a study of a correlation between skin color and IQ. *Intelligence*, *34*(2), 131–137. https://doi.org/10.1016/j.intell.2005.04.004

Iliescu, D., Ilie, A., Ispas, D., Dobrean, A., & Clinciu, A. I. (2016). Sex differences in intelligence: A multi-measure approach using nationally representative samples from Romania. *Intelligence*, *58*, 54–61. https://doi.org/10.1016/j.intell.2016.06.007

IMF. (2024). *GDP per capita*. https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD

Jensen, A. R. (1970). A theory of primary and secondary familial mental retardation. *International Review of Research in Mental Retardation*, *4*, 33–105. https://doi.org/10.1016/s0074-7750(08)60022-1

Jensen, A. R. (1980). *Bias in mental testing*. Free Press.

Kamin, L. J. (2006). African IQ and mental retardation. *South African Journal of Psychology*, *36*(1), 1–9. https://doi.org/10.1177/008124630603600101

Karwowski, M., & Milerski, B. (2021). Intensive schooling and cognitive ability: A case of Polish educational reform. *Personality and Individual Differences*, *183*, 111121. https://doi.org/10.1016/j.paid.2021.111121

Kirkegaard, E. O. W. (2021). An examination of the openpsychometrics.org vocabulary test. *OpenPsych*. https://doi.org/10.26775/op.2021.07.05

Kirkegaard, E. O. W., & Karlin, A. (2020). National intelligence is more important for explaining country well-being than time preference and other measured non-cognitive traits. *Mankind Quarterly*, *61*(2), 339–370. https://doi.org/10.46469/mq.2020.61.2.11

Lasker, J., & Kirkegaard, E. O. W. (2022). *The generality of educational effects on cognitive ability: A replication*. Center for Open Science. http://dx.doi.org/10.31234/osf.io/8s2vx

Lee, J. J., McGue, M., Iacono, W. G., Michael, A. M., & Chabris, C. F. (2019). The causal influence of

brain size on human intelligence: Evidence from within-family phenotypic associations and GWAS modeling. *Intelligence*, *75*, 48–58. https://doi.org/10.1016/j.intell.2019.01.011

Lerner, B. (1983). Test Scores as Measures of Human Capital. In *Intelligence and National Achievement* (pp. 65–99). Institute for the Study of Man.

Lurija, A. R. (1978). *Cognitive development: Its cultural and social foundations*.

Lynn, R. (1978). Ethnic and racial differences in intelligence: International comparisons. In *human variation: The biopsychology of age, race, and sex*. Academic Press.

Lynn, R. (2006). *Race differences in intelligence: An evolutionary analysis*. Washington Summit Publishers.

Lynn, R., & Cheng, H. (2013). Differences in intelligence across thirty-one regions of China and their economic and demographic correlates. *Intelligence*, *41*(5), 553–559. https://doi.org/10.1016/j.intell.2013.07.009

Lynn, R., & Meisenberg, G. (2010). The average IQ of sub-Saharan Africans: Comments on Wicherts, Dolan, and van der Maas. *Intelligence*, *38*(1), 21–29. https://doi.org/10.1016/j.intell.2009.09.009

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Praeger.

Lynn, R., & Vanhanen, T. (2012). National IQs: A review of their educational, cognitive, economic, political, demographic, sociological, epidemiological, geographic and climatic correlates. *Intelligence*, *40*(2), 226–234. https://doi.org/10.1016/j.intell.2011.11.004

Odell, B., Gierl, M., & Cutumisu, M. (2021). Testing measurement invariance of PISA 2015 mathematics, science, and ICT scales using the alignment method. *Studies in Educational Evaluation*, *68*, 100965. https://doi.org/10.1016/j.stueduc.2020.100965

OECD. (2022). *PISA 2022 technical report*. OECD. https://www.oecd.org/publications/pisa-2022-technical-report-01820d6d-en.htm

Patrinos, H. A., & Angrist, N. (2018). *Global dataset on education quality: A review and update (2000–2017)*. World Bank, Washington, DC. http://dx.doi.org/10.1596/1813-9450-8592

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence*, *41*(6), 791–801. https://doi.org/10.1016/j.intell.2013.06.005

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains. *Perspectives on Psychological Science*, *10*(3), 282–306. https://doi.org/10.1177/1745691615577701

Pineau, J.-C., Delamarche, P., & Bozinovic, S. (2005). Les Alpes Dinariques : Un peuple de sujets de grande taille. *Comptes Rendus. Biologies*, *328*(9), 841–846. https://doi.org/10.1016/j.crvi.2005.07.004

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Recueil, C. (2023). With the latest PISA results, America has proven once again that it has one of the smartest populations and, perhaps, the best education systems. *Crémieux*. https://twitter.com/cremieuxrecueil/status/1732087511327908128

Recueil, C. (2024). Final Preview: Here are age-stratified Flynn effects before and after correcting for bias. *@cremieuxrecueil*. https://x.com/cremieuxrecueil/status/1749258628022796790

Reichenberg, A., Cederlöf, M., McMillan, A., Trzaskowski, M., Kapra, O., Fruchter, E., Ginat, K., Davidson, M., Weiser, M., Larsson, H., Plomin, R., & Lichtenstein, P. (2015). Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proceedings of the National Academy of Sciences*, *113*(4), 1098–1103. https://doi.org/10.1073/pnas.1508093112

Rindermann, H. (2018). *Cognitive capitalism: Human capital and the wellbeing of nations*.

Rindermann, H., & Becker, D. (2018). FLynn-effect and economic growth: Do national increases in intelligence lead to increases in GDP? *Intelligence*, *69*, 87–93. https://doi.org/10.1016/j.intell.2018.05.001

Ritchie, S. J., Bates, T. C., & Deary, I. J. (2015). Is education associated with improvements in general cognitive ability, or in specific skills? *Developmental Psychology*, *51*(5), 573–582. https://doi.org/10.1037/a0038981

Rushton, J. P. (1997). Cranial size and IQ in Asian Americans from birth to age seven. *Intelligence*, *25*(1), 7–20. https://doi.org/10.1016/s0160-2896(97)90004-0

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*(2), 235–294. https://doi.org/10.1037/1076-8971.11.2.235

Saß, S., Kampa, N., & Köller, O. (2017). The interplay of g and mathematical abilities in large-scale assessments across grades. *Intelligence*, *63*, 33–44. https://doi.org/10.1016/j.intell.2017.05.001

Sear, R. (2022). *'National IQ' datasets do not provide accurate, unbiased or comparable measures of cognitive ability worldwide*. Center for Open Science. http://dx.doi.org/10.31234/osf.io/26vfb

Shinwari, A., Véron, A., Abdianwall, M. H., Jouve, E., & Laporte, R. (2022). Tap water consumption is associated with schoolchildren's cognitive deficits in Afghanistan. *International Journal of Environmental Research and Public Health*, *19*(14), 8252. https://doi.org/10.3390/ijerph19148252

Templer, D. I., & Arikawa, H. (2006). Temperature, skin color, per capita income, and IQ: An international perspective. *Intelligence*, *34*(2), 121–139. https://doi.org/10.1016/j.intell.2005.04.002

Travaglino, P., Meazza, C., Pagani, S., Biddeci, G., & Bozzola, M. (2011). Secular trends in growth of African Pygmies and Bantu. *HORMONES*, *10*(2), 144–148. https://doi.org/10.14310/horm.2002.1304

United Nations. (2023). *Human development index*. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Warne, R. T. (2022). National mean IQ estimates: Validity, data quality, and recommendations. *Evolutionary Psychological Science*, *9*(2), 197–223. https://doi.org/10.1007/s40806-022-00351-y

Warne, R. T. (2023). *Stupid? No. Unfamiliar? Yes. The meaning of low mean IQs in developing nations – Russell T. Warne*. https://russellwarne.com/2023/04/29/stupid-no-unfamiliar-yes-the-meaning-of-low-mean-iqs-in-developing-nations/

Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (2010). *WAIS-IV clinical use and interpretation: Scientist-Practitioner perspectives*. Academic Press.

Weiss, L. G., Saklofske, D. H., Holdnack, J. A., & Prifitera, A. (2019). *WISC-V assessment and interpretation: Clinical use and interpretation*. Academic Press.

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & van der Maas, H. L. J. (2010). Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn Effect. *Learning and Individual Differences*, *20*(3), 135–151. https://doi.org/10.1016/j.lindif.2009.12.001

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span,

M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509–537. https://doi.org/10.1016/j.intell.2004.07.002

Wikipedia. (2024). *Programme for international student assessment*. Wikipedia. https://en.wikipedia.org/wiki/Programme_for_International_Student_Assessment

World Bank Open Data. (2023). *GDP per capita*. World Bank Open Data. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

World Economics. (2023). *Population data quality ratings methodology*. World Economics. https://www.worldeconomics.com/Concepts/Population-Data-Quality-Ratings/

Wu, A., Li, Z., & Zumbo, B. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment, Research, and Evaluation*, *12*(3), 1–26.