Evolutionary Psychological Science
https://doi.org/10.1007/s40806-022-00351-y

RESEARCH ARTICLE

National Mean IQ Estimates: Validity, Data Quality, and Recommendations

Russell T. Warne[1]

Received: 30 September 2022 / Revised: 27 November 2022 / Accepted: 28 November 2022

Abstract

Estimates of mean IQ scores for different nations have engendered controversy since their first publication in 2002. While some researchers have used these mean scores to identify relationships between the scores and other national-level variables (e.g., economic and health variables) or test theories, others have argued that the scores are without merit and that any study using them is inherently and irredeemably flawed. The purpose of this article is to evaluate the quality of estimates of mean national IQs, discuss the validity of different interpretations and uses of the scores, point out shortcomings of the dataset, and suggest solutions that can compensate for the deficiencies in the data underpinning the estimated mean national IQ scores. My hope is that the scientific community can chart a middle course and reject the false dichotomy of either accepting the scores without reservation or rejecting the entire dataset out of hand.

Keywords International differences · IQ · National IQs · Validity

In 2002, Lynn and Vanhanen (2002) published a dataset of estimated mean IQ scores for 185 nations and found a correlation between those mean scores and per capita income. In the 20 years that have elapsed since those estimated mean scores were published, the reception of national IQs has varied from wholesale dismissal (e.g., Ebbesen, 2020; Hunt & Sternberg, 2006; Kamin, 2006; Sear, 2022) to total acceptance and use of the unaltered scores for later data analyses (e.g., Belasen & Hafer, 2013; Kanazawa, 2006; Piffer, 2013; Templar & Arikawa, 2006)

In the ensuing decades, the national IQ estimates have been updated several times (Lynn & Meisenberg, 2010; Lynn & Vanhanen, 2006, 2012), with the current version of the dataset of national mean IQs being updated by Lynn and Becker (2019a) and available online for download at https://viewoniq.org (Lynn & Becker, 2019b). This most recent iteration of the database adds data, corrects errors, and eliminates data from poorly documented or incorrectly included samples. As a result, many of the criticisms of earlier versions of the database because of the inclusion of specific samples (e.g., Dickins et al., 2007; Kamin, 2006; Wicherts et al., 2010a, c, d) are less applicable today because many of the most criticized samples are no longer included in the most recent version of Lynn and Becker's (2019b) dataset.

Lynn and Becker (2019a, b) also expanded their dataset to include national-level economic, health, educational, geographic, and other data. Lynn and Becker (2019a) used much of these non-cognitive data to demonstrate that their estimated mean IQs correlated with a wide variety of national-level variables and had utility in understanding international differences in many aspects of societal development.

Although the national IQ dataset has improved since its first release, the controversy surrounding the data has not diminished. One recent article using the dataset caused an uproar and was retracted shortly thereafter (see Bauer, 2020, for the editor's statement). Another article that used an earlier version of the national IQ dataset (Nyborg, 2012) was the subject of an investigation in which the author was found to have committed scientific dishonesty. (The verdict later was overturned in court, and the panel was ordered to pay a judgment to the author.) On social media and blogs, there are regular calls from scientists and activists to retract papers based on the national IQ dataset, and work using the scores is often denounced as immoral or racist (e.g., Ebbesen, 2020).

The purpose of this article is to examine the Lynn and Becker (2019b) dataset and evaluate the validity of evidence surrounding their mean IQ estimates for nations. This is not

✉ Russell T. Warne
russwarne@gmail.com

[1] Provo, Utah, USA

Published online: 19 December 2022

an apologetic piece; I believe that there are weaknesses as well as strengths in the national IQ dataset. In writing this article, my hope is that the scientific community can chart a middle course and reject the false dichotomy of either accepting the scores without reservation or rejecting the entire dataset out of hand.

A Note About IQ

Before discussing the Lynn and Becker (2019b) dataset, it is important to briefly discuss the nature of IQ. The term is often misunderstood because of incorrect beliefs that many people have about intelligence (Warne, 2020; Warne & Burton, 2020) and because of historical and cultural baggage surrounding the term "IQ." The first vital point to understand is that IQ—whether a score for an individual or a mean for a group—is merely a measurement or number and is not the same as the underlying construct of intelligence (see also Haier, 2017; Hunt & Carlson, 2007; Warne, 2020; and Wicherts et al., 2010a, c; for this distinction). Consequently, I ask readers to not think of "IQ" and "intelligence" as always being synonymous. While in some populations the IQ score produced by an intelligence test may be validly interpreted as corresponding to an examinee's intelligence level, this is not always the case. In this article, I use the term "IQ" as a convenience, mostly to follow Lynn and his colleagues' own use and because the term is more concise than alternatives like "cognitive test performance." Readers should not assume that I am interpreting an average IQ from Lynn and Becker's (2019b) dataset as measuring a country's mean intelligence level. Indeed, in some parts of this article, I will explicitly state that such an interpretation is sometimes not justified.

The second important point is that IQ is a measure of a phenotype. There is no assumption of innateness in a reported IQ score. Indeed, this is true for all psychometric test scores, including measures of personality, mental health, and opinions. Test scores cannot measure a person's genetic or inborn potential. The same applies to groups, including nations; finding that one group has a higher mean score than another group is merely a statement of current average phenotypes. Such mean differences can be due to a variety of causes, including testing artifacts (see Warne et al., 2014), educational differences (Ritchie & Tucker-Drob, 2018), or a lack of experience with the test stimuli (e.g., Serpell & Jere-Folotiya, 2008). Additional data are required to determine why one group has a higher average score than another. The national IQ dataset cannot—by itself—provide explanations for international differences in IQ, nor do its compilers claim that it does. Likewise, national IQs are not set in stone; many researchers believe that these values can or will change over time (Rindermann et al., 2017), including Lynn and Becker (2019a, Chapter 4) themselves.

Can National Mean IQs Even be Calculated?

Before evaluating Lynn and Becker's (2019b) dataset, it is legitimate to ask whether the endeavor of calculating the mean IQ of a nation is theoretically or scientifically tenable. In other words, can national mean IQs even be calculated? There are three reasons why the answer is yes.

First, summary statistics, such as means, can be calculated for any interval-level variable for any human population. Calculating means for a wide variety of measures—including IQ scores—is a common practice in the social and biological sciences. Estimating a group's average is a reasonable practice, and there is no reason why this practice cannot be extended to IQ scores.

Second, national averages are acceptable, as shown in data from international organizations that report national averages for age, per capita income, and other characteristics. National averages have also been calculated for psychological and behavioral variables, such as personality scores (Terracciano et al., 2005), educational test performance (Patel & Sandefur, 2020), and crime rates (United Nations Office on Drugs & Crime, 2022). Again, there is no logical reason why IQ scores cannot be included in the list of variables for which national averages can be calculated.

Finally, there is the question of whether IQ scores have the statistical characteristics that produce undistorted averages. In other words, it is important to determine whether IQ scores are interval-level data and therefore suitable for calculating averages, or whether they are ordinal-level data that preclude the calculation of means. Whether test scores are ordinal- or interval-level data is a debate that dates to the earliest days of this data classification scheme (Stevens, 1946). In a brief survey of this debate, Warne (2021, pp. 29–30) found that some scholars have argued that test scores are always ordinal data, while others claim that they are interval-level data. The debate may never be fully settled, but for IQ scores, the evidence is strong that they are interval-level data. The test scores function statistically as expected from interval data (Jensen, 1969, 1998), and deviations from this performance are rare. Therefore, calculating mean IQ scores does not violate any statistical assumptions.

Can IQs from Different Samples be Combined?

The main underlying data in the Lynn and Becker (2019b) dataset are mean IQ values for scores from multiple samples that were then combined to form an overall mean estimated IQ for the entire country. This methodology raises questions about whether it is scientifically tenable to combine mean IQs from different samples to find a single overall mean. Just as with the last section, there are three reasons to believe that combining samples in this way is justifiable.

First, combining means from different samples into one overall mean is no different than conducting a meta-analysis to combine summary statistics (e.g., effect sizes, correlation coefficients) from different samples into an overall value. From this perspective, Lynn and his colleagues merely conducted a series of small meta-analyses (one for each country) in order to identify an average IQ for the entire population in each country. Given the long history of meta-analysis (dating back to Glass, 1976) and its widespread acceptance in psychology, medicine, and other fields (Williams et al., 2017), Lynn and his colleagues' decision to statistically combine datasets' mean IQs to produce a national mean IQ is tenable, and their work should be judged by the same standards by which any meta-analysis is judged.

Second, there is a question of whether different tests are measuring the same underlying construct within a country, which would be a prerequisite to combining mean scores on different tests to calculate an overall national average IQ. Tests that produce an IQ score supposedly measure a global problem-solving ability that humans use on a wide variety of cognitive tasks. The best evidence that this is so comes from a study of data from 31 non-Western, economically developing countries. In exploratory factor analysis procedures, a single underlying factor emerged in 94 of 97 (96.9%) samples, indicating that an overall reasoning ability "... appears in many cultures and is likely a universal phenomenon in humans" (Warne & Burningham, 2019, p. 237). Moreover, the average variance explained by this underlying factor was 45.9%, which is similar to results from Western samples. While the authors of this study cautioned that it does not permit cross-national comparisons (Warne & Burningham, 2019, p. 266), their work does show that within nations, multiple samples almost always show a single cognitive factor. This gives researchers confidence that the procedure of combining IQs from different samples collected in the same country is an empirically supported practice because the tests can measure the same underlying ability.

Another theoretical challenge to the procedure of combining IQ scores from different tests into a national average is the fact that tasks on intelligence tests are often very disparate (Jensen, 1980). Indeed, some psychologists scrutinizing the content of different intelligence tests have questioned whether the tests really measure the same construct (e.g., Helms, 1992; Richardson, 2002). The reality is that the surface appearance of an intelligence test item matters little. As long as a task or item on a test requires cognitive effort from examinees, it will measure intelligence to some extent. This principle is called the "indifference of the indicator," and it was identified nearly a century ago (Spearman, 1927). Because of the indifference of the indicator, the underlying factor scores produced by different tests correlate .89 or higher (Floyd et al., 2013, pp. 389–396; Johnson et al., 2004, p. 95; Johnson et al., 2008, p. 88; Keith et al., 2001, p. 108; Stauffer et al., 1996, p. 193). The indifference of the indicator is so pervasive that some researchers accidentally created intelligence tests when trying to measure other abilities (Warne, 2020). As a result, researchers can combine different IQs from different tests into an overall mean with little concern that test-specific characteristics will distort the final result.

Background of the National Mean IQ Estimates

Lynn and Vanhanen (2002) originally developed the national mean IQs to examine whether there was a relationship between mean IQ and a nation's per capita income. It is a reasonable scholarly question to investigate whether population characteristics influence economic growth or development. Economists since Adam Smith have asked the same question (Angrist et al., 2021), and they continue to do so today (e.g., Clark, 2007; Gust et al., 2022; Jones, 2016), and some social scientists have used national IQ estimates for the same purpose (e.g., Rindermann, 2018a). Moreover, there is a positive correlation at the individual level between IQ and income (Murray, 2002; Zisman & Ganzach, 2022), which means that calculating mean IQ estimates to investigate this question at the national level is a reasonable scientific endeavor.

Lynn and his coauthors updated the dataset multiple times (Lynn & Meisenberg, 2010; Lynn & Vanhanen, 2006, 2012). The current version was mostly updated by David Becker, and he and Lynn described the methodology in a book (Lynn & Becker, 2019a) and in the documentation that accompanies the dataset itself. Lynn and Becker's (2019a) work was aimed at increasing the level of detail in the description of the methodology so that readers have easy access to all the important information about underlying samples.

Some of the improvements that Lynn and Becker (2019a) made were global changes that impacted many national mean IQ estimates in their dataset. For example, they refined their equations used to convert scores on matrix tests to IQ scores and improved their adjustments for the Flynn effect[1] (i.e., the steady increase of IQ scores seen worldwide in the twentieth century). Lynn and Becker also included ratings of sample quality, test quality, and the calculation procedures for every sample.

[1] This Flynn effect adjustment is often misunderstood. It does not increase or decrease the score of the country to reflect the age of the test. Rather, it adjusts the international IQ standard (where 100 = the mean in the UK) to the year of the test administration in a country so that the country's measured IQ is compared to the estimated standard for the same year.

Other improvements are more basic. For example, in earlier versions of the dataset, some source data were poorly sourced or included incomplete citations, a problem that was replicated in other work that was based on the mean national IQs (e.g., Lynn, 2015). By rechecking every score used in earlier versions of the database, Lynn and Becker (2019a) have improved the quality of the estimated national mean IQs.

At a detailed level, Lynn and Becker's (2019a) documentation includes an explanation of every dataset and the procedures that led to the estimate to a mean IQ for a sample. This allows users to look up any country and understand the data that contributed to its mean IQ score and how that mean was calculated. This is especially important for datasets that present unique situations that require subjective decisions to handle. The amount of documentation in Lynn and Becker (2019a) and online is impressive and shows a commitment to transparency. Few meta-analyses can boast this level of detail, and the information bolsters confidence in Lynn and Becker's procedures and results.

Briefly, the methodology for Lynn and Becker's (2019b) latest national IQ estimates is as follows. First, they would identify a qualifying sample from a country and record a mean score on an intelligence test and ancillary information (e.g., sample age, sample size, year and location of data collection, and special sample characteristics). In this step, any subgroups within a sample (e.g., different ages, sex, and geographic groups) were recorded separately. If needed, the mean test score was converted to the standard IQ metric (with a mean of 100 and a SD of 15). Afterward, a correction for the Flynn effect was applied so that samples that received older tests did not have a systematic advantage over samples that took newer tests. The next step was to correct for the country that the test was normed in, which was necessary because if IQ differences exist between countries where samples were normed, then this would introduce a positive bias in samples using tests normed in countries with lower mean IQs (and a negative bias in samples using tests normed in countries with higher mean IQs). The UK was arbitrarily chosen as the country that other norming countries' data would be adjusted to.

Lynn and Becker (2019a, b) also calculated IQs from educational achievement tests that were administered to representative samples of school children in different countries. The tests used for this purpose were the Progress in International Reading Literacy Study (PIRLS), Program for International Student Assessment (PISA), and Trends in International Mathematics and Science Study (TIMSS). For countries where these tests were administered, scores were collected, and if a test was administered to the same grade or age group in multiple years, the scores for that test and group were averaged. These scores were then converted to a *z*-score using the mean and standard deviation for corresponding students in the UK, and this *z*-score was then converted to an IQ score with a mean of 100 and a standard deviation of 15.

These procedures produced a mean IQ for 149 countries in Lynn and Becker's (2019b) current dataset. For the remaining 52 countries that lacked educational or intelligence test data, the authors used a geographic imputation procedure. To calculate these imputations, Lynn and Becker (2019a, b) took advantage of the spatial autocorrelation that often exists in international data and identified the three countries with the longest land borders that had IQ means. A mean IQ, weighted by the length of the land border, was calculated and used as an estimate for the country's missing estimated mean IQ. For island nations, the three closest countries with IQ data were identified, and an unweighted mean was calculated and imputed as an estimated mean IQ value for the missing country's data.

At the end of these procedures, Lynn and Becker (2019a, b) reported up to six mean IQ estimates for a country:

1. An unweighted mean of all adjusted IQs for each sample within a country (UW IQ),
2. A mean of all adjusted IQs weighted by the size of each sample within a country (NW IQ),
3. A mean of all adjusted IQs weighted by the size and quality of each sample within a country (QNW IQ),
4. A mean IQ based on educational achievement data from the TIMSS, PIRLS, or PISA tests (SAS IQ),
5. An unweighted mean of the previous two estimated IQs (QNW + SAS IQ), and
6. A mean equal to either the QNW + SAS IQ or a mean imputed from neighboring countries that do have IQ data (QNW + SAS + GEO IQ).

For any country with more than one recorded sample IQ, these means vary slightly. The average difference between the minimum and maximum mean IQs for countries (excluding those with no variation because they are based on one sample or source of data) is 5.77 points (median = 4.48 points, SD = 5.06 points) in the most recent version of the national IQ dataset (Lynn & Becker, 2019b). However, there are a few outliers in this respect; the largest discrepancies are found in Egypt and Vietnam, which have over 20 IQ points between their highest and lowest estimated IQs. For both of these countries—and for almost all others with at least a 7-point discrepancy—this is caused by the SAS IQ being substantially lower or higher than the other estimated IQs for the country. Because of this within-country variation, it is always important for researchers who use these national IQs to specify which mean they are using and to justify their selection.

Validity and National IQs

Nature of Validity

There has been debate about the meaning and utility of national IQ scores. Often, this debate takes the form of discussions of "the validity of national IQs" (Lynn & Meisenberg, 2010, p. 353). This is an unfortunate perspective because, according to the *Standards for Educational and Psychological Testing*, "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA] et al., 2014, p. 11). In other words, validity is not a property of scores themselves, but rather validity is a property of score *uses* or *interpretations*. Therefore, it is incorrect to refer to "the validity of national IQs." Indeed, referring to "the validity of scores" or "test validity" is inaccurate for any psychometric score or test (AERA et al., 2014, p. 23)—not just national IQs.

A test score can be valid for one use or interpretation and invalid for another. For example, an end-of-year educational achievement score can often be validly interpreted as a measure of what a student has learned over the course of the school year. But interpreting that same score to be a measure of teacher quality or school excellence often is not empirically justified (Warne, 2020, p. 76). The former interpretation is valid, while the latter is not—even though the scores are the same.

Moreover, validity must be considered anew for each novel use, interpretation, or examinee population for a test (AERA et al., 2014, Chapter 1). It cannot automatically be assumed that a valid interpretation or use of a test for one population will apply to a new interpretation, use, or population. Extending a test in this fashion requires gathering new validity evidence (AERA et al., 2014, Standards 1.2 & 1.4). For this reason, sweeping generalizations that national mean IQs are valid or invalid are hasty oversimplifications. Anyone encountering these claims should immediately ask *which* interpretations, uses, and/or populations that national IQs are valid (or invalid) for.

Data Quality in the National IQ Dataset

As important as validity is, it is inherently limited by the quality of the underlying data. One of the most pervasive criticisms of Lynn and Becker (2019b) dataset—and its predecessors—is the quality of the data (e.g., Ebbesen, 2020; Sear, 2022; Wicherts et al., 2010b, d). Undeniably, the quality of data varies across countries. Combined sample sizes (not including samples from educational tests) range from 19 (for Angola) to 62,649 (for the USA), for example. Likewise, some countries' estimated IQs are based on much more representative samples than others'. And 52 nations have IQs estimated from neighboring countries because they have no data of their own.

Lynn and Becker (2019a, b) recognized this heterogeneity and provided ratings of quality for every sample. These quality ratings are coded from descriptions of the original articles and fall into three categories: test characteristics, sample characteristics, and IQ calculation procedures. A coding system was developed that allows characteristics to have between 0 and 1 points. Within each category, these scores are summed, and then the three categories are equally weighted to produce an overall quality rating. The coding scheme is briefly described below.

- Quality ratings based on test characteristics:
  - If an entire test was administered (i.e., all items or all subtests), then the sample receives 1 quality point. Samples administered a portion of a test (e.g., an abbreviated version or a limited number of subtests) receive 0 quality points.
  - Samples that were administered a test less than 10 years after the test's standardization received 1 quality point. If a test was administered 10 to 20 years after standardization, then the sample received 0.5 points. Tests administered more than 20 years after standardization received 0 points.
  - If a sample took a test that was standardized in the UK, then it received 1 quality point. Otherwise, the sample received 0 quality points.
- Quality ratings of the sample characteristics:
  - Data collected from representative groups samples are given the highest score (1 point), followed by samples taken from a specific region, one or more rural areas, or one or more urban areas (0.5 points); samples of people living in a different country (i.e., immigrants or refugees tested outside of their birth country) received 0 quality points.
  - Samples that are considered to have a typical socioeconomic status for the target population are given the most weight (1 point). Samples of wealthier or poorer individuals received 0 points.
  - Samples composed of normal individuals or twins receive the highest score (1 point). All other samples received 0 quality points.
  - Samples characterized as representative samples and samples that are norm groups for psychometric tests received a quality of 1 point. Samples described as random received 0.5 points. All other samples—including convenience samples—received 0 quality points.
  - If a sample's mean age deviates by less than 10 years from the median age of the country, then the sample is

awarded 1 quality point. Samples with a mean age that is 10 to 20 years away from the national median receive 0.5 quality points. All other samples received 0 points.

- Quality ratings based on the calculations needed to convert the test score to the international IQ metric:
  - Samples that reported scores that were either in the IQ metric or raw scores that could be converted to IQs via procedures described in the test manual or with the aid of the test norms received 1 quality point. Otherwise, 0 quality points were awarded.
  - If a sample had a mean score that needed no corrections for testing artifacts or procedures (e.g., time limits, converting the sample's test score to the score for another test), then the sample received 1 point in its quality rating. All other samples received 0 quality points.
  - If there were no special corrections needed to convert the reported score to the international IQ metric, then the sample received 1 point for its quality rating. If special corrections were required, then the sample received 0 quality points.

In theory, these ratings ranged from 0 to 1, though in practice, the observed range was 0.18 to .90 points.[2] The mean sample quality rating is .5653 (median = .62, SD = .1663). This quality rating system shows that Lynn and Becker (2019b) already took into consideration sample representativeness (in terms of geography, socioeconomic status, age, normality), test characteristics, and the uncertainty introduced when converting scores to a different test or scale. The QNW IQ is an average that weights for sample size and these characteristics of data quality, which makes it superior to the UW and NW IQs that Lynn and Becker (2019b) also provide.

Finding the correlation between sample quality and IQ is informative for gauging the degree of systematic bias that IQ may have as a consequence of data quality. Sample mean IQ and sample quality ratings are uncorrelated ($r = -.004$), indicating that the quality of a sample's data is unrelated to the mean IQ for that sample. At the national level, the mean sample quality and UW IQ are $r = -.197$. Together, these correlations indicate that sample quality is not systematically biasing IQ estimates at the sample level and that, at the national level, lower IQ estimates for countries have a slight tendency to be based on data of *higher* mean quality, contrary to the expectation of the critics of the national IQ dataset (e.g., Ebbesen, 2020).

The sample size criticism is much less of a problem. The minimum sample size to have a 95% confidence interval width of ± 3 IQ points is 96. For a confidence interval of ± 2 IQ points, a combined sample size of 216 is needed for a country. And countries with mean IQs based on combined samples of 856 or more will have a confidence interval that does not exceed ± 1 IQ point around the mean.[3] Setting aside the geographically-based estimates for countries that have no IQ data, the average combined sample size for a national-level IQ estimate is 4730.9 (median = 2018; SD = 8631.7). The sample size in most countries far exceeds that needed to produce a narrow confidence interval around an estimated population mean. How wide a confidence interval should be in order to have a desired level of exactitude is a subjective decision (Warne, 2021, pp. 199–201), but of 131 countries with UW, NW, or QNW IQs in the Lynn and Becker (2019b) dataset, only five (Angola, the Democratic Republic of the Congo, the Dominican Republic, Greenland, and Uzbekistan) have a 95% confidence interval wider than ± 3 IQ points.

With 683 mean IQ scores (not counting educational achievement testing data) contributing to Lynn and Becker's (2019b) estimated national IQs, it is easy to cherry-pick a few low-quality samples. The question is not whether low-quality samples contributed to IQ data; everyone agrees that they did, and low-quality data are not unusual in meta-analyses. However, there seems no compelling reason to believe that these low-quality samples systematically bias the IQ estimates or are so numerous that the entire dataset should be thrown out.

Correspondence to Alternative National IQ Calculations

Any meta-analysis requires scientists to make subjective decisions, and these decisions can have an important influence on the final result. One way to check the results of a meta-analysis is to perform the same analysis and determine whether the results are similar to the original. For the Lynn and Becker (2019b) dataset, this requires identifying the correlations between their national mean IQ estimates and the estimates from other datasets. If the correlations are high, then it is unlikely that subjective decisions are biasing the results for the scores.

Table 1 reports the intercorrelations of 16 different measures of national IQ and reported (along with all other data for original analyses reported here) in this article's Supplemental File. The first six are the UW IQ, NW IQ, QNW IQ, QNW + SAS IQ, and

[2] Only one sample had an overall quality rating of .18; it was collected in the United States. Four samples achieved an overall quality rating of .90. The data for these samples were collected in Tajikistan, the UK, the USA, and Yemen.

[3] The width of a confidence interval is equal to $\pm 1.96(\frac{\sigma}{\sqrt{n}})$, where $\frac{\sigma}{\sqrt{n}}$ is equal to the standard error of the mean, $\sigma = 15$ (the default SD of a population on the IQ metric), and n is the combined sample size of all samples that contribute to a country's mean IQ estimate (Warne, 2021, pp. 199–201).

Table 1 Correlations among IQs and educational assessment score from worldwide or multiregional reports

Data source	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1. Lynn and Becker (2019b) UW	1.000	131	131	82	131	131	0	73	124	111	130	130	129	62	112	113
2. Lynn and Becker (2019b) NW	.972	1.000	131	82	131	131	0	73	124	111	130	130	129	62	112	113
3. Lynn and Becker (2019b) QNW	.969	.999	1.000	82	131	131	0	73	124	111	130	130	129	62	112	113
4. Lynn and Becker (2019b) SAS	.792	.775	.777	1.000	100	100	0	55	95	82	100	100	100	64	98	95
5. Lynn and Becker (2019b) QNW + SAS	.948	.968	.969	.964	1.000	149	0	74	139	115	148	148	147	70	130	130
6. Lynn and Becker (2019b) QNW + SAS + GEO	.948	.968	.969	.964	1.000	1.000	52	80	181	133	197	197	194	78	158	155
7. Lynn and Becker (2019b) GEO	–[a]	–[a]	–[a]	–[a]	–[a]	1.000	1.000	6	42	18	49	49	47	8	28	25
8. Lynn and Vanhanen (2002)	.901	.885	.885	.904	.903	.909	.976	1.000	81	78	81	81	81	46	73	74
9. Lynn and Vanhanen (2002) + GEO	.872	.855	.853	.878	.856	.870	.872	1.000	1.000	125	183	183	183	78	151	152
10. Lynn and Vanhanen (2012)	.902	.899	.897	.915	.919	.919	.845	.971	.957	1.000	133	133	132	67	113	113
11. Lynn and Vanhanen (2012) + GEO	.871	.860	.856	.915	.880	.874	.801	.970	.946	1.000	1.000	199	196	78	157	157
12. Becker and Rindermann (2016)	.873	.856	.852	.961	.894	.879	.781	.957	.938	.980	.978	1.000	196	78	157	157
13. Rindermann (2018b, pp. 18–22)	.869	.851	.847	.970	.894	.881	.793	.961	.940	.974	.973	.995	1.000	78	156	156
14. Patel and Sandefur (2020)	.790	.787	.787	.941	.896	.871	-.376	.851	.794	.847	.838	.894	.881	1.000	77	77
15. Angrist et al. (2021) HLO	.814	.813	.811	.966	.871	.829	.140	.900	.841	.888	.866	.911	.896	.938	1.000	145
16. Gust et al. (2022)	.826	.814	.812	.936	.868	.865	.556	.887	.868	.919	.901	.911	.913	.885	.931	1.000

Numbers in the lower half of the table are correlations. Numbers in the upper half of the triangle are the number of countries used to calculate the correlation. Pairwise deletion was used in the calculation of all correlations

[a]No correlation because countries only had a GEO IQ if there were no other data available to calculate the UW, NW, QNW, SAS, and QNW + SAS IQs

QNW + SAS + GEO IQs reported in Lynn and Becker's (2019b) dataset. The next measure, the GEO IQ, is derived from the data file and consists solely of the 52 geographically imputed IQ scores for countries that have no cognitive or educational test data. The following four scores are national IQs reported from earlier versions of Lynn's work (Lynn & Vanhanen, 2002, 2012). This is followed by IQs reported by Becker and Rindermann (2016) and Rindermann (2018b), which are partially based on Lynn and Vanhanen's (2012) IQs, supplemented with data from international educational achievement tests. The final three scores listed in Table 1 are reported in studies used to calculate measures of educational achievement across multiple continents or worldwide (Angrist et al., 2021; Gust et al., 2022; Patel & Sandefur, 2020). It is important to note that none of the authors who produced data in the final three rows of the correlation table made claims that they were estimating national intelligence averages.

For five of Lynn and Becker's (2019b) IQ scores, the intercorrelations are very high: UW IQ, NW IQ, QNW IQ, QNW + SAS IQ, and QNW + SAS + GEO IQ all intercorrelate $r = .948$ to 1.000, as shown in Table 1. This is unsurprising because all of these are based on the underlying data that overlap greatly, though it does show that decisions made at the data combination or later calculation stages of analysis make little difference in the results. The outlier in Lynn and Becker's (2019b) scores are the SAS IQs, which correlate $r = .797$ with UW IQ, $r = .775$ with NW IQ, $r = .777$ with QNW IQ, and $r = .964$ with QNW + SAS IQ and QNW + SAS + GEO. This aligns with the finding that countries with differences between the minimum and maximum IQs of more than 7 points are almost always due to a discrepancy between the SAS IQ and the others.

Another method of checking Lynn and Becker's (2019b) IQs is to compare the scores to previous calculations from Lynn and Vanhanen (2002, 2012). Table 1 shows that the correlations between the Lynn and Becker (2019b) national IQs and the national IQs from prior versions of the dataset are very high: $r = .853$ to .909 for the first Lynn and Vanhanen (2002) dataset and $r = .856$ to .919 for a later version (Lynn & Vanhanen, 2012). This is in spite of the fact that Lynn and Becker (2019a) re-evaluated every dataset and recalculated their scores from the original data (though many of the underlying samples are the same).

The intra-country changes in IQ from Lynn and Vanhanen (2012) to Lynn and Becker (2019b) are also informative when evaluating the degree to which subjective decisions can impact the results. The IQ changes from the old database to the new one's QNW + SAS + GEO IQ were generally small: an average of 4.40 IQ points (median = 3.01 IQ points, SD = 4.48 IQ points),[4] indicating that any future recalculation of national IQs based on the same data would likely produce similar results. However, it is important to note that large discrepancies between the two datasets did occur for some countries. Eighteen nations[5] had a change of 10 IQ points or more; usually, these countries had sparse data, which made their IQs more susceptible to changes when data were added or deleted by Lynn and Becker (2019b).

Finally, there are two non-Lynn sources for national IQs, though they do draw on his work in the calculations. First, there is another calculation from Becker and Rindermann (2016), which was based on Lynn and Vanhanen's (2012) scores, but with educational achievement testing data included. These correlated $r = .852$ to .961 with Lynn and Becker (2019b) estimated mean national IQ scores. A few years later, Rindermann (2018a) calculated new national IQ estimates, partially basing his work on Lynn and Vanhanen's (2012) scores. Rindermann's (2018a) corrected IQ values correlate $r = .776$ to .970 with the six IQs from Lynn and Becker (2019b).

Given these high correlations, it is unlikely that subjective decisions have made a substantial impact on the Lynn and Becker (2019b) dataset in general. If one takes these correlations as a measure of reliability (e.g., interrater reliability), then the IQs are generally consistent enough for research purposes. However, individual countries—especially when their estimated average IQ is based on a small number of samples or the scores of neighboring countries—can show substantial fluctuations from one calculation to another.

External Validity Evidence

External validity evidence is data showing that a psychometric score correlates with other variables (called *criteria*) that are not part of the test. External validity evidence is essential because a test score that does not correlate with any criteria has no meaning outside of the testing situation. Such a test score is essentially useless. Therefore, identifying nonzero correlations between national IQ means and other national-level data is an essential prerequisite for any argument that there are any valid interpretations or uses for national IQs (AERA et al., 2014, pp. 17–18, 28–30).

Lynn and Vanhanen (2002) have always recognized this need, and in the first publication on the mean national IQ

[4] These statistics are calculated using the absolute value of the differences between the QNW + SAS + GEO IQ in the Lynn and Becker (2019b) dataset and the IQ + GEO IQ for the previous version.

[5] Listed in descending order of the magnitude of IQ change: Nicaragua (−23.78 IQ points); Haiti (21.60 IQ points); Honduras (−18.84 IQ points); Nepal (−18.00 IQ points); Guatemala (−17.71 IQ points); Saint Helena, Ascension, and Tristan da Cunha (−17.01 IQ points); Belize (−16.25 IQ points); Cabo Verde (−16.00 IQ points); Morocco (−15.39 IQ points); Yemen (−14.39 IQ points); Mauritania (−14.00 IQ points); Chad (11.83 IQ points); Saint Lucia (11.71 IQ points); Barbados (11.69 IQ points); Senegal (−10.50 IQ points); Republic of the Congo (−10.03 IQ points); Côte d'Ivoire (−10.02 IQ points); and Vanuatu (10.02 IQ points). Positive values in this list indicate that the new IQ estimates from Lynn and Becker (2019b) are higher than the earlier estimate. Negative values indicate the new value is lower.

estimates, they published data on the correlates of national IQs. Over the years, Lynn and Becker (2019a) have continued to accumulate this validity evidence, and now the latest volume devotes over 100 pages to discussing correlations with national IQs. It is infeasible to describe all this information in detail; however, I do believe it is beneficial to describe some highlights to show that these mean national IQ estimates do correlate with other national-level variables that are theoretically expected to be related to a nation's average test performance.

Economic Characteristics

At the individual level, income and IQ are positively correlated in Western nations (Murray, 2002; Zagorsky, 2007; Zisman & Ganzach, 2022). Therefore, it is reasonable to expect a positive correlation between the two variables at the group level. To test this supposition, Lynn and Becker (2019a, pp. 218–240) have compiled dozens of variables that correlate with their mean national IQ estimates. Although exact correlations vary, depending on the choice of IQ and economic measure, the results are very consistent that the two variables are correlated; nations with a higher mean IQ are wealthier both in the twenty-first century and in the past, have stronger long-term economic growth, and experience less income inequality. Some local conditions may weaken or even temporarily eliminate these correlations, such as an abundance of natural resources, a strong tourism-based economy, a history of communism, and price shocks. In general, however, the pattern of national IQ correlating with favorable economic conditions holds (Lynn & Becker, 2019a).[6]

Educational Achievement

Intelligence tests were first created in an educational context (Wolf, 1973), and they are still frequently used for educational purposes over 100 years later. At the individual level, the evidence is strong and consistent that IQ positively correlates with educational achievement (Jensen, 1998; Warne, 2020). In fact, IQ is the single best predictor of educational outcomes (Deary et al., 2004; Roth et al., 2015; Zaboski et al., 2018; Zhao et al., 2019). This makes it reasonable to expect a positive correlation between IQ and beneficial educational outcomes at the group level.

Three of Lynn and Becker's (2019b) IQ scores use educational achievement tests as a data source, which makes positive correlations between these IQs and educational achievement variables unsurprising. Still, other research has shown that international achievement scores measure a global cognitive adeptness and that subject-specific academic content is a minor source of variation (Pokropek et al., 2022). This supports Lynn and his colleagues' decisions to use educational achievement tests as a source of data for estimating mean national IQ.

Just as they did for economic data, Lynn and Becker (2019a, pp. 202–215) gathered dozens of educational variables in order to examine correlations with their estimated mean national IQ scores. The results are consistent in showing that nations with higher estimated IQs are more educated, have better educational systems (both in terms of inputs—such as higher expenditures and better teachers—and outputs, like high school graduation rates), and perform better on educational achievement tests.

Non-educational Outputs

School is not an end to itself; nations educate children and young adults in order to prepare them for participation in society and the economy. Therefore, it is legitimate to ask whether IQ scores correlate with life variables outside of the schoolhouse. There is abundant evidence that at the individual level, they do. High-IQ samples of individuals identified as children or adolescents earn higher incomes, have more prestigious jobs, produce more creative works (e.g., patents, books, scientific articles), and have a higher quality of life than the general population (Bernstein et al., 2019; Holahan & Sears, 1995; Lubinski et al., 2014; Makel et al., 2016; Terman & Oden, 1947, 1959). This relationship is monotonic, with no apparent IQ threshold where the relationship between IQ and these outcomes weakens or reverses (Kell et al., 2013; Lubinski, 2009; Wai, 2014).

Lynn and Becker (2019a, pp. 240–245, 281–282, 293–295) gathered data that show that beneficial non-educational outcomes at the national level are also positively correlated with national mean IQ estimates. Per-capita scholarly papers, patents, Nobel Prizes, innovation, transportation safety, and similar outcomes are higher in countries with higher estimated mean IQs. Conversely, unfavorable societal outputs are lower in countries that have higher estimated mean IQs, including deforestation, pollution, crime, and corruption. These correlations are evidence that the national IQ values are measuring something that is related to people's daily lives.

Health, Mortality, and Wellness

Another consistent finding at the individual level is that higher-IQ people tend to enjoy better physical health (Hart et al., 2004; Sörberg et al, 2014) and have fewer mental health problems (Gale et al., 2010; Woodberry et al., 2008). Smarter people also tend to live longer (Beaver et al., 2016; Deary et al., 2004; Hart et al., 2003), and a negative correlation

[6] This finding also occurs in cross-national comparisons of educational achievement test scores. See, for example, Angrist et al. (2021), Gust et al. (2022), and Patel and Sandefur (2020).

between mortality and IQ has been shown at the county level within the USA (Barnes et al., 2013).

Lynn and Becker (2019a, pp. 266–280) reported a similar range of health variables that correlated with their estimates of the national mean IQ. The national IQs correlate positively with life expectancy and negatively with mortality at all age levels, malnutrition, and infectious disease rates. This is exactly the pattern one would expect, based on the correlations between health-related variables and IQ at the individual level.

Political Institutions

The final set of variables that are relevant to the issue of the external validity data regarding national IQs is related to nations' political institutions. Lynn and Becker (2019a, pp. 245–253) presented evidence that estimated IQs for nations correlate positively with demographic government, economic freedom, the rule of law, government effectiveness, political rights, protection of intellectual and physical property, and the quality of political institutions. Conversely, estimated national IQs are negatively correlated with corruption, gender inequality, and war. This is unsurprising, given the positive correlation at the individual level between good decision-making and IQ (see Jones, 2016, for an in-depth discussion of this research).

Discussion

My sole purpose in describing these correlations is to show that the national IQs are measuring an important population characteristic. Countries with higher estimated IQs are generally more prosperous, better educated, more innovative, healthier, and more democratic. In short, whatever these national IQs measure, it is clear that higher scores generally appear in countries that make their citizens' lives better, and these benefits even seem to extend beyond a country's borders (Rindermann & Carl, 2020). For that reason alone, these scores should not be dismissed out of hand.

Readers should remember the old cliché that "correlation is not causation." I do not wish to imply that the national IQ level has any causal impact on any of these country-level outcomes, though some have made that claim (e.g., Kanazawa, 2006; Rindermann, 2018a). Indeed, no causal explanation for a correlation with a test score is necessary to establish evidence for the validity of an interpretation or use of that score. Rather, my goal is to show that national IQs have abundantly met one of the essential requirements—nonzero correlations with external criteria—for there to be valid interpretations or uses of these scores.

Convergent Validity Evidence

Another form of external validity evidence comes from convergent validity data, which consist of strong correlations between two tests that supposedly measure the same construct. For national IQ estimates, there have been a number of international studies using intelligence tests or educational achievement tests that can serve to provide convergent validity evidence of IQs within a region. In these studies, researchers have used data that are independent of Lynn and Becker (2019b) to calculate national IQs or to produce scores that can be easily converted to national IQs.

Latin America

One notable regional study that can be used to check Lynn and Becker's (2019b) national IQ estimates was the Study of Latin American Intelligence (SLATINT) project (Flores-Mendoza et al., 2018). Using the Raven's Standard Progressive matrices in six large cities in Argentina, Brazil, Chile, Colombia, Mexico, and Peru, Flores-Mendoza et al. (2018, p. 95) estimated the average IQs in these samples to be between 86 and 94. These estimates are higher than Lynn and Becker's (2019b) QNW + SAS IQs for these countries, though this is at least partially due to the data coming from urban areas, which tend to have higher IQ averages than rural areas (e.g., Zhao et al., 2019). For five of the six countries, the difference is 6 IQ points or less, and the correlation between the IQ values in these countries is $r = .842$. Flores-Mendoza et al. (2018, p. 107) also calculated an estimated IQ of 97 for Spain, which is just 3.13 points higher than the QNW + SAS IQ of 93.87 in the Lynn and Becker (2019b) data.

In SLATINT, the outlier is the Peruvian sample, which had a mean IQ of 94, a value that Flores-Mendoza et al. (2018, p. 97) called "unexpectedly high." The value of 94 is almost a full standard deviation higher than the estimate of Lynn and Becker's (2019b) QNW + SAS IQ estimate of 81.42 based on nine samples (total $n = 2702$). After briefly reviewing the literature on intelligence testing in Peru, Flores-Mendoza et al. (2018, pp. 97–98) concluded that their sample overestimated Peruvian IQ. The Peruvian sample was the least representative sample in SLATINT, with 92% of students attending private schools and 63% in high socioeconomic status schools. The parents were also extremely well educated, with 82% of fathers and 94% of mothers having attended college (Flores-Mendoza et al., 2018, p. 32). Including the Peruvian sample in the analysis reduces the correlation between SLANTINT IQs and Lynn and Becker's (2019b) QNW + SAS IQs to $r = .129$.

Nineteen countries in Latin America have participated at least once in educational tests administered by the Laboratorio Latinoamericano Evaluación Calidad Educación (LLECE), a division of the United Nations Educational, Scientific and Cultural Organization (UNESCO). In chronological order, these tests were called PERCE 1997, SERCE 2006, TERCE 2013, and ERCE 2019. LLECE administers educational achievement tests to students in two elementary grades (third and fourth grades for PERCE 1997 and third

and sixth grades for the other tests) in each nation. These tests can be used to create a composite score (like the SAS IQ from Lynn & Becker, 2019b) that can then be used to rank order the countries according to their overall educational achievement.

Data from the LLECE tests are available online.[7] I accessed these data, and every national mean scale score was converted to a *z*-score using the mean and standard deviation of the combined international sample for the same year, grade, and school subject. After transforming every national mean scale score into a national *z*-score, I followed the same procedure that Lynn and Becker (2019a, pp. 39–40) used to calculate a SAS score from international educational achievement test data. In short, I averaged a nation's *z*-scores within a given year and then averaged these to form a composite *z*-score for that nation. The correlation between the composite *z*-score from the LLECE tests and Lynn and Becker's (2019b) QNW + SAS IQ is $r = .547$.

This correlation seems low, especially compared to the intercorrelations between student achievement scores and estimated national IQ scores within Lynn and Becker's (2019b) dataset. However, the data from one outlier country—Cuba—is distorting these correlations. The PERCE 1997 and SERCE 2006 mean scores for Cuba are unrealistically high: $z = +1.86$ to $+2.02$ for PERCE 1997 scores and $z = +.96$ to $+1.48$ for the SERCE 2006 data (when comparing each test score to the regional mean across all countries). These means were all 1.30 to 2.00 standard deviations above the second-highest scoring country on the PERCE 1997 and 0.33 to 1.58 standard deviations higher than the second-highest scoring country on the SERCE 2006.[8] Rindermann (2018b) suggested that the Cuban data are outliers because totalitarian countries may tamper with the data that produce international rankings to raise their prestige. When the outlier of Cuba was removed, the correlation increased to $r = .633$.

The regional validation from Flores-Mendoza et al. (2018) and the LLECE tests provide a substantial degree of convergent validity evidence indicating that Lynn and Becker's (2019b) national IQ estimates and LLECE academic achievement scores measure the same underlying construct. Although both sets of criteria data have an outlier distorting the results, these outliers are easily explained, and eliminating them increased the positive correlations between IQ estimates and the data from the other studies.

[7] The PERCE 1997 and SERCE 2006 data are taken from official publications reporting country means for each grade level and subject (Oficina Regional de Educación para América Latina y el Caribe/UNESCO, 2001, p. 176; 2008, Tables A.3.1, A.3.5, A.4.1, A.4.5, and A.5.1). TERCE 2013 and ERCE 2019 data can be downloaded at https://raw.githubusercontent.com/llece/comparativo/main/datos_grafico_1-1.csv

[8] Cuba did not participate in TERCE 2013. Its ERCE 2019 data are much more similar to data from other countries in Latin America.

Sub-Saharan Africa

By far, the most controversy regarding national IQs has been in regard to the accuracy of national IQs in sub-Saharan Africa. In every version of Lynn and his colleagues' national IQs, sub-Saharan African countries have had some of the lowest estimated mean IQs. Gathering validity evidence for these countries is particularly important because these countries often have the sparsest data; if there is a satisfactory level of validity evidence about the use and interpretation of national IQs in sub-Saharan Africa, then it instills confidence in the entire dataset.

The best validity data regarding Lynn and Becker's (2019b) IQs comes from the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) examinations. This is a regional educational achievement test administered in 1995, 2000, and 2007 in 6 to 14 southern and eastern African countries. QNW + SAS IQ scores correlate $r = .851$, .754, and .861 with the 1995, 2000, and 2007 SACMEQ scores, respectively. The QNW + SAS + GEO IQ scores correlate $r = .866$, .726, and .806 with the 1995, 2000, and 2007 SACMEQ scores, respectively.[9]

Another testing program that provides evidence regarding convergent validity in sub-Saharan Africa is the Monitoring Learning Achievement (MLA) Project, sponsored by UNESCO in the 1990s. The tests administered for the MLA Project were not as sophisticated as the SACMEQ tests. Results for 11 African countries in the 1999 test administration were only reported as a percentage of questions answered correctly on three different tests: life skills, literacy, and numeracy (Chinapah et al., 2000, pp. 21, 84–107). Of the three tests, numeracy correlated most highly with Lynn and Becker's (2019b) QNW + SAS IQ ($r = .745$) and QNW + SAS + GEO IQ ($r = .657$). Results were weaker for the literacy ($r = .411$ for QNW + SAS IQ and $r = .428$ for QNW + SAS + GEO IQ) and life skills ($r = .199$ for QNW + SAS and $r = .445$ for QNW + SAS + GEO) tests. When the scores for the three tests were combined, the unweighted MLA mean score correlated $r = .609$ with QNW + SAS IQ and $r = .606$ with QNW + SAS + GEO IQ.

Another source of enlightening African educational achievement data is the tests administered in the Programme d'Analyse des Systèmes Éducatifs de la Confemen (PASEC), an international educational achievement testing program that is administered in 10 francophone countries in western and central Africa. The 2014 PASEC tests were administered in grades 2 and 6, with each grade receiving a language and mathematics exam. Of the 10 countries participating in the 2014

[9] The 1995 SACMEQ test only produced reading scores. The 2000 and 2007 SACMEQ tests produced a reading and mathematics score. The 2000 and 2007 scores were combined as an unweighted mean for each country when calculating correlations with the estimated national-level IQs.

PASEC tests (Benin, Burkina Faso, Burundi, Cameroon, Chad, Congo, Côte d'Ivoire, Niger, Senegal, and Togo), only Benin and Burkina Faso have received an IQ score from Lynn and Becker (2019b) based on IQ data from samples in those countries. The other eight nations only have an IQ score imputed from neighboring countries. Finding correlations between PASEC scores and the QNW + SAS IQ scores for Benin and Burkina Faso is an uninformative analysis because a correlation for a sample size of 2 has zero degrees of freedom (Warne, 2021, Formula 12.7) and will always result in a correlation of $r = \pm 1$. However, all 10 nations have a QNW + SAS + GEO IQ score, and the correlations between these scores and the PASEC academic achievement scores are not impressive: for Grade 2 PASEC scores, the correlation is $r = .164$ for language and $r = .013$ for mathematics. For Grade 6 PASEC scores, the correlation with QNW + SAS + GEO IQ is $r = -.461$ for language and $r = -.110$ for mathematics. This shows that geographically imputed scores have a poor correspondence with data drawn from a country.[10]

Multiregional Validation

Regional validation data for national estimated mean IQ scores are useful, but the number of countries is limited (20 or less), making these correlations unstable due to outliers, the restriction of range, and the small sample size. A multiregional effort to rank order nations in their cognitive test performance provides a better gauge of the accuracy of Lynn and Becker's (2019b) national IQ scores. The most common international educational achievement tests are the PIRLS, PISA, and TIMSS tests. However, these scores contribute data to Lynn and Becker's (2019b) SAS, QNW + SAS, and QNW + SAS + GEO IQ scores, and any positive correlations between the PIRLS, PISA, and TIMSS scores on the one hand and the SAS, QNW + SAS, and QNW + SAS + GEO IQ scores on the other will be inflated because of the overlap in data.

The best multiregional ranking of nations' educational achievement that is not based entirely on PISA, PIRLS, and TIMSS that I have been able to find is in a study by Patel and Sandefur (2020). They constructed a new test using publicly available items from the TIMSS, PIRLS, PASEC, and LLECE tests and administered it to a representative sample of students in a low-scoring region of India. They then used published parameters for these items (based on the item response theory models that are used to score these tests) and their sample's performance to link the new test to the other four, thereby equating the scales so they could convert scores from LLECE and PASEC to the PIRLS and TIMSS scale. Patel and Sandefur (2020) released reading and mathematics PIRLS/TIMSS scores for 79 countries on all six inhabited continents.[11]

I converted these PIRLS/TIMSS scores to a national IQ by calculating a *z*-score (using the American mean and the international standard deviation) for each country and then converting this to an IQ by multiplying the result by 15 and adding 100. I then subtracted 2.5 IQ points to center this score on Lynn and Becker's (2019a, b) IQ scale on the mean for the UK.[12] The reading and mathematics scores were then averaged to produce an estimated overall IQ for each country. This new IQ correlated $r = +.896$ with Lynn and Becker's (2019b) QNW + SAS IQ and $r = .871$ with their QNW + SAS + GEO IQ, as shown in Table 1.

However, discrepancies between the IQs that I derived from Patel and Sandefur's (2020) data and Lynn and Becker's (2019b) estimated IQs were apparent. The eight countries that only had a geographically imputed IQ score (Burundi, Cameroon, Chad, Côte d'Ivoire, Niger, Paraguay, Senegal, and Togo) in Lynn and Becker's (2019b) dataset had a correlation of $r = -.376$ between the two IQs (see Table 1), showing again that the geographically imputed IQs did not correspond with data from the countries.

Another source of data for multiregional validation is the Programme for the International Assessment of Adult Competencies (PIAAC). This testing program, created by the Organization for Economic Cooperation and Development (OECD), differs from other international assessments in that the examinees are not school children, but rather working-age individuals. The 2012 and 2015 PIAAC tests were administered to representative samples of individuals, aged 16–65, in 34 nations and regions.[13] Europe and North America are heavily represented in the PIAAC assessment, but countries in East Asia (Indonesia, Japan, Singapore, South Korea), Central Asia (Russia), and Latin America (Chile) also participated.

[10] The correlations between PASEC scores and the GEO IQ scores from the Lynn and Becker (2019b) dataset—i.e., with Benin and Burkina Faso removed—are $r = .142$ (for PASEC grade 2 language), $r = .153$ (for PASEC grade 2 mathematics), $r = -.662$ (for PASEC grade 6 language), and $r = -.262$ (for PASEC grade 6 mathematics). This does not change the conclusion that geographically imputed scores have a poor correspondence with data drawn from a country.

[11] The national PIRLS/TIMSS scores and the chart to convert LLECE and PASEC scores to PIRLS/TIMSS scores to one another are available at https://www.cgdev.org/sites/default/files/patel-sandefur-human-capital-final-results.xlsx

[12] I used the American mean in this calculation because the UK was not one of the countries in Patel and Sandefur's (2020) study. The 2.5 IQ point adjustment is the standard adjustment that Lynn and Becker (2019b) used when examinees took a test normed in the USA instead of the UK.

[13] Two regions, England and Northern Ireland, were part of the same country. When calculating correlations with QNW + SAS IQs, the Northern Ireland data were dropped, and the data for England was compared to QNW + SAS IQs for the entire UK.

The PIAAC tests measure mastery of literacy and numeracy on a common scale. Using data taken from the 2012 and 2015 PIAAC tests (OECD, 2016, Fig. 2.22), the correlation between Lynn and Becker's (2019b) QNW + SAS IQs is $r = .760$ for the literacy assessment, $r = .708$ for the numeracy assessment, and $r = .744$ for the mean of the two PIAAC scores. These high correlations are an important piece of convergent validity evidence regarding Lynn and Becker's (2019b) international IQ estimates because the examinees are all 16 years of age or older. In Lynn and Becker's (2019a, b) data, though, the mean age for a country's data when calculating UW, QW, or QNW IQs exceeded 18 years in only 28 of 131 countries (21.4%). The high correlations between the QNW + SAS IQs and the PIAAC scores show that calculating scores for a country based on data from children does not invalidate the use of Lynn and Becker's (2019b) IQs, as Sear (2022) has claimed. This is an unsurprising finding because of data at the individual level that show (1) high test–retest correlations of IQ scores across the lifespan (Deary et al., 2013; Starr et al., 2000) and (2) a strong correlation between academic achievement and IQ (Deary et al., 2004, 2007; Zaboski et al., 2018; Zhao et al., 2019). This also supports the widespread use among economists of data from academic assessments of children in order to make interpretations about an entire nation's human capital (e.g., Angrist et al., 2021; Gust et al., 2022; Hanushek, 2016).[14]

Worldwide Validation

Even better than convergent validity evidence from a multiregional study would be a worldwide independent study that could provide convergent validity evidence regarding Lynn and Becker's (2019a, 2019b) estimated mean national IQs. There are two[15] new worldwide datasets—both based on international academic achievement data—that can be used to check Lynn and Becker's (2019b) national IQ estimates. These datasets were created by Angrist et al. (2021) and Gust et al. (2022).

In 2021, Angrist et al. (2021) published the Human Learning Outcomes (HLO) database,[16] which is affiliated with the World Bank and based on international educational achievement test scores, including large multinational exams like PIRLS, PISA, TIMSS, and regional exams, like SACMEQ, LLECE, and PASEC. These scores are harmonized to be placed on the same metric so that nearly every nation in the world can be compared with one another. The only inclusion requirement is that a country has participated in at least one international educational achievement assessment since 2000. The HLO database covers 164 nations.

As I did with the Patel and Sandefur (2020) data, I converted every HLO score into a *z*-score, and then averaged a country's different *z*-scores within each year and then averaged the resulting *z*-scores across multiple years to produce a single *z*-score for each country. These values were then converted to IQ scores, referenced to the mean of the UK and using the international scale's standard deviation.

The correlation between these IQs derived from the HLO data replicated the regional and multiregional analyses in this article. Table 1 shows that the correlation between HLO IQs and Lynn and Becker's (2019b) QNW IQs was $r = .811$, even though the two IQs are based on completely independent data. The correlation between HLO IQs and the SAS IQs was $r = .966$, which is partial because the scores are based on some of the same data (i.e., PIRLS, PISA, and TIMSS scores). Still, this large correlation shows that subjective decisions—whether made by the HLO team or Lynn and Becker—are not a major source of error in SAS IQs or HLO scores. The correlation between HLO IQs is $r = .896$ with QNW + SAS IQs and $r = .830$ with QWN + SAS + GEO IQs, though it is important to keep in mind that these are not fully independent because of some shared underlying data. Finally, the correlation between HLO-derived IQs and the geographically imputed IQs in Lynn and Becker's (2019a, b) dataset was $r = .140$, another indication that geographically imputed IQs often have a poor correspondence with data drawn from those countries.

In 2022, Gust and et al. used PISA, TIMSS, LLECE, PARSEC, and SACMEQ test data to harmonize combined math and science achievement combined scores on a single scale for 126 countries. They estimated via regression scores for 33 additional countries using economic development and school enrollment data from other countries in the geographic region.

Comparing Gust et al.'s (2022) scores to Lynn and Becker's (2019b) national mean IQ estimates produces results that are similar to those seen when comparing IQs to the Angrist et al. (2021) HLO data. Table 1 shows that Gust et al.'s (2022) scores correlate $r = .812$ with Lynn and Becker's (2019b) QNW IQs,

[14] Sear's (2022) criticism of using IQ data from children to estimate IQs for an entire population shows that she does not understand that IQ scores are calculated by comparing examinees to their age peers. This functionally controls for age and allows scores from different age groups to have the same meaning. For an accessible explanation of how IQ scores are calculated, see Warne (2020), pp. 5–9.

[15] Readers may be aware of Lim et al.'s (2018) study that measures human capital in 195 countries. These scores are not included in the discussion in this article because the underlying data are not solely cognitive/educational scores. Lim et al. (2018) also used health data and longevity/life expectancy data in the calculation of their human capital scores. Therefore, the Lim et al. (2018) data cannot be interpreted as a cognitive measure, which makes it inadequate to use for convergent validity purposes when studying the Lynn and Becker (2019b) dataset.

[16] Available at https://datacatalog.worldbank.org/search/dataset/0038001.

even though the two sets of scores originate from different data (i.e., international achievement scores vs. scores on cognitive tests) and were derived via different procedures. When correlating Gust et al.'s (2022) national scores with the SAS IQs, the correlation rises to $r = .936$, which is expected because of the overlapping sources of data used to calculate both scores. The correlation between Gust et al.'s (2022) national scores on the PISA scale correlates $r = .868$ with Lynn and Becker's (2019b) QNW + SAS IQs and $r = .865$ with the QNW + SAS + GEO IQs (see Table 1). These correlations are all very similar to the correlations with the Angrist et al. (2021) data, even though the three scores were calculated by different teams using different statistical procedures and different (though overlapping) data. Such strong correlations show that, again, the subjective decisions used to create the Lynn and Becker (2019b) data did not have a strong influence on the results.

Plausibility of Extremely Low National IQs

The correlational data between Lynn and Becker's (2019b) national IQ estimates and other measures of cognitive competence is helpful in establishing an agreement in the rank order of countries in multiple datasets, whether on a regional, multiregional, or worldwide scale. However, many critics of Lynn and his colleagues' work (e.g., Ebbesen, 2020; Kamin, 2006; Sear, 2022) find the IQs in some nations implausibly low and thereby question the usefulness of the entire dataset. Correlations address this criticism only partially because it is statistically possible that there could be a consistent downward bias of scores in economically developed nations, even with a high correlation. Additionally, because the mean and the variance of any dataset are statistically independent,[17] investigating means, in addition to correlations (which are variance-based statistics) would strengthen the validity of evidence regarding the national IQ estimates.

Most of the criticism about low IQs in the Lynn datasets is directed at the low IQs calculated for many sub-Saharan African countries (Ebbesen, 2020; Kamin, 2006; Sear, 2022; Wicherts et al., 2010a, b, c, d), and the bulk of estimated low estimated mean QNW + SAS IQs are in this region of the world.[18] Many of the same scores that showed high correlations with Lynn and Becker's (2019b) IQs can be converted to the same IQ scale in order to examine whether Lynn and Becker's estimates are reasonable.

Low IQs from Other Researchers

The HLO-derived IQs (from Angrist et al., 2021) show that very low IQs are possible when independently calculated without reference to Lynn and colleagues' work. In the HLO database, 43 of 164 (26.2%) countries received a score that was the equivalent of an IQ less than 75. Half of these countries (22 of all 164 nations, or 13.4%) score below 70. Angrist et al. (2021) defined 300 on their scale as being "low performance," which is 2.13 standard deviations below the mean in the UK and the equivalent of an IQ of 68.11. For 17 countries, the average performance on an educational test was at or below this level. The lowest score—obtained by Sierra Leone— is the equivalent of an IQ of 58.81.

As Angrist et al. (2021) did, Gust et al. (2022) found very low scores compared to the UK mean. Converting Gust et al.'s (2022) scores to the international IQ scale using the mean and standard deviation[19] for the UK produces IQs that range from 55.0 (for Niger) to 108.6 (for Singapore). Across all 159 countries, 39 (24.5%) have a score that is the equivalent of an IQ of 75 or lower—which is almost the exact same percentage found in the Angrist et al. (2021) data (26.2%). Of these, 31 had an IQ below 70 (31 of all 159 countries, or 19.5%). This shows, again, that very low IQs can be derived independently of the methods of Lynn and Becker (2019a, b).

Where the results depart from the Angrist et al. (2021) results is in the correlation between Gust et al.'s (2022) scores and the geographically imputed IQs for 25 countries that do not have cognitive data in Lynn and Becker (2019b). The correlation between these two scores is $r = .556$, which is much higher than the correlation with the Angrist et al. (2021) scores ($r = .140$). It is not clear why the correlation with geographically imputed IQs is so much higher in the Gust et al. (2022) dataset.[20] This is the only evidence supporting the use of Lynn and Becker's (2019b) geographically imputed scores.

[17] This statistical truism is why the Flynn effect (a purely environmental effect) can coexist with high heritability (a variance statistic measuring the strength of generic influence on a phenotype in a population) of IQ. The same secular mean increase occurred in height (a phenotype with high heritability) in many countries during the twentieth century. Changes in the mean do not automatically result in changes in the variance—and vice versa.

[18] It is important to recognize that mean QNW + SAS IQs below 70 are also found in some Central American nations (Belize, El Salvador, Guatemala, Honduras, Nicaragua), the Caribbean (Dominica and Saint Vincent and the Grenadines), and Morocco, Nepal, and Yemen.

[19] For the 2018 PISA, the SD for the UK data was 93 for math scores and 99 for science scores (Schleicher, 2019, pp. 7–8). In these calculations, I used the standard deviation of 99 to be more conservative. My choice of standard deviation will not affect any correlations, but it will change differences between these IQs and others and make outlier national mean IQs slightly less extreme.

[20] This is not an artifact of the extrapolation based on nearby countries' data that Gust et al. (2022) used. The correlation between scores for the 12 countries that had imputed data in both datasets was $r = .608$; for the 13 countries that had geographically imputed scores in the Lynn and Becker (2019b) dataset and scores based on educational achievement testing data in the Gust et al. (2022) dataset, the correlation was $r = .511$. The average difference between the two sets of scores is also similar.

Table 2 Regional and worldwide mean IQs (weighted by population)

IQ measure	Mean (No. of countries)								
	World Bank region[a]								
	Central Asia	East Asia and Pacific	Europe	Latin America and Caribbean	Middle East and North Africa	North America	South Asia	Sub-Saharan Africa	World
Lynn and Becker (2019b) UW	91.2 (5)	98.6 (17)	94.5 (35)	85.8 (22)	79.7 (19)	92.6 (3)	75.4 (5)	70.4 (25)	86.6 (131)
Lynn and Becker (2019b) NW	90.9 (5)	99.5 (17)	93.7 (35)	85.8 (22)	80.6 (19)	95.8 (3)	78.0 (5)	70.2 (25)	87.7 (131)
Lynn and Becker (2019b) QNW	91.4 (5)	99.0 (17)	94.1 (35)	86.0 (22)	80.7 (19)	96.1 (3)	78.3 (5)	70.4 (25)	87.7 (131)
Lynn and Becker (2019b) SAS	98.0 (3)	98.2 (15)	95.4 (42)	81.6 (15)	73.2 (18)	99.2 (2)	73.6 (1)	60.5 (4)	88.6 (100)
Lynn and Becker (2019b) QNW + SAS	93.8 (5)	98.6 (18)	94.6 (44)	83.1 (28)	77.5 (21)	97.6 (3)	76.3 (5)	69.7 (25)	86.7 (149)
Lynn and Becker (2019b) QNW + SAS + GEO	93.6 (6)	98.4 (33)	94.6 (45)	83.2 (38)	77.5 (21)	97.6 (3)	76.4 (8)	69.4 (46)	86.2 (201)
Lynn and Becker (2019b) GEO	85.5 (1)	93.4 (15)	95.3 (1)	83.7 (10)	–[b]	–[b]	82.3 (3)	68.4 (21)	76.6 (52)
Lynn and Vanhanen (2002)	96.0 (1)	98.0 (16)	97.9 (25)	87.6 (13)	84.5 (7)	97.9 (2)	80.9 (2)	68.3 (15)	89.3 (81)
Lynn and Vanhanen (2002) + GEO	93.6 (6)	97.6 (29)	97.4 (40)	87.2 (31)	84.3 (20)	97.9 (2)	81.0 (8)	68.9 (47)	88.0 (183)
Lynn and Vanhanen (2012)	95.3 (2)	100.9 (22)	96.1 (32)	86.0 (24)	83.9 (19)	97.8 (3)	82.2 (5)	70.7 (26)	89.7 (133)
Lynn and Vanhanen (2012) + GEO	91.5 (6)	100.5 (33)	95.7 (45)	85.5 (35)	83.9 (21)	97.8 (3)	82.1 (8)	70.8 (47)	89.1 (199)
Becker and Rindermann (2016)	92.8 (6)	98.4 (33)	95.9 (45)	83.6 (35)	82.7 (21)	98.9 (3)	77.4 (8)	70.6 (47)	87.1 (199)
Rindermann (2018b, pp. 18–22)	92.7 (6)	96.9 (32)	95.8 (45)	83.8 (33)	82.9 (21)	98.3 (3)	78.7 (8)	70.2 (47)	86.9 (196)
Patel and Sandefur (2020)	97.4 (2)	89.1 (8)	91.6 (26)	76.8 (16)	71.8 (12)	97.4 (2)	—[b]	67.8 (12)	85.6 (78)
Angrist et al. (2021) HLO	98.6 (4)	100.5 (25)	97.3 (43)	84.5 (23)	79.0 (19)	100.8 (2)	74.7 (6)	69.8 (41)	87.3 (158)
Gust et al. (2022)	92.3 (6)	100.7 (21)	96.7 (43)	83.9 (21)	81.9 (18)	98.4 (2)	73.3 (6)	69.9 (41)	86.9 (157)

Means are calculated by weighting by national population (taken from Lynn & Becker, 2019b). Only countries with at least one IQ value in the Lynn and Becker (2019b) dataset are included in these calculations

[a]Geographic regions were defined by the World Bank, with the exception of splitting the Europe and Central Asia region, following Gust et al.'s (2022, p. 16) division

[b]No IQ estimates available in the region

Regional Means

Table 2 reports 16 estimates of average IQ scores (weighted by population) for different regions of the world. Thirteen of these are the work of Lynn and his colleagues or are based partially on their work. However, the final three rows (Angrist et al., 2021; Gust et al., 2022; Patel & Sandefur, 2020) are IQs derived from authors working independently and basing their calculations solely on international educational achievement data and provide an important check on Lynn's work.[21] Readers should note, though, that these authors use data that overlaps with the data source for Lynn and Becker's (2019a, b) SAS, QNW + SAS, and QNW + SAS + GEO IQs.

Comparing the average IQs for different regions shows that Lynn and Becker's (2019b) values are in line with previous versions of the dataset and with the estimates derived from studies published by other teams. Regional mean IQs based on Angrist et al.'s (2021) HLOs for different world regions range from 69.7 for sub-Saharan Africa to 100.8 for North America. All the regional weighted averages for

[21] Gust et al. (2022, p. A1) noted that Angrist et al.'s (2021) method overestimates academic achievement HLOs, compared to the Gust et al. (2022) method. The average scores in Table 2 are much more similar than would be expected because of the different means for the UK that were used to calculate *z*-scores and IQs. The HLO mean for the UK

Footnote 22 (continued)

is 527.8 in the Angrist et al. (2021) data, compared to the Gust et al. (2022) mean of 503.2. The higher HLO mean for the UK provides a correction to the HLO scores, when converted to IQs, and makes the weighted mean IQs for both datasets in Table 2 much more similar.

the HLO-derived IQs are similar to the QNW + SAS IQs that Lynn and Becker (2019b) calculated. The largest difference is for Central Asia (5.1 IQ points higher HLO IQs than QNW + SAS IQs), and the smallest discrepancy is 0 points, which was found by both groups of researchers for sub-Saharan Africa. The worldwide weighted population average difference between the two groups is just 0.6 IQ points, with the HLO IQs slightly higher. The regional average IQs derived from the Gust et al. (2022) data show a very similar pattern. The largest difference is in the Middle East and North Africa region (with Gust et al., 2022, IQs being 4.4 IQ points higher), and the smallest difference in sub-Saharan Africa (where Lynn & Becker's, 2019a, b, QNW + SAS IQ for the region is 0.2 points lower). These comparisons show that Lynn and his colleagues' IQs for different regions of the world can be corroborated by other teams of researchers working independently.

Most notably, the regional mean IQs derived from Patel and Sandefur's (2020) data were slightly *lower* than almost all of Lynn and Becker's (2019b) QNW + SAS IQs, with the only exception being Central Asia, which had a mean of 3.6 IQ points higher in Patel and Sandefur's (2020) data. Patel and Sandefur's (2020) data produced a mean IQ for sub-Saharan Africa of 67.8 IQ points, further evidence that other researchers can produce findings that indicate extremely poor performance on psychometric tests in this region.

The comparisons in Table 2 must be made with caution because different IQ means were calculated with different collections of the country within a region. An informative comparison is to examine the difference between IQs derived for the same country by Patel and Sandefur (2020) and Lynn and Becker (2019b) independently. When comparing each individual nation, the IQs derived from Patel and Sandefur (2020) are an average of 4.42 IQ points lower than the QNW + SAS IQs and 2.62 IQ points lower than the QNW + SAS + GEO IQs. This is an important finding for two reasons. First, it is another sign that the extremely low IQs that Lynn and Becker (2019b) estimate for sub-Saharan Africa are likely not systematically underestimating African IQs, as earlier critics have claimed (e.g., Sear, 2022; Wicherts et al., 2010c, d), nor are these scores implausibly low, as is the opinion of some more recent critics (e.g., Ebbesen, 2020; Sear, 2022).

Sub-Saharan African IQ and Educational Achievement Data

Whether IQs for some African countries are unrealistically low can be tested by comparing estimated mean IQs with the international educational achievement data that are available for some of these nations (Wicherts et al., 2010c). There are three sub-Saharan African nations that have taken part in the PIRLS, PISA, or TIMSS testing: Botswana, Ghana, and South Africa. Lynn and Becker (2019b) used pre-2019 data to calculate SAS IQs for these countries. All three nations have scores that are below the UK average on TIMSS and PIRLS: 1.27 to 2.67 SDs lower for Botswana, 1.76 to 2.78 SDs lower for Ghana, and 1.26 to 2.95 SDs lower for South Africa. On the IQ scale, these values range from 55.75 to 80.95. In comparison, the SAS IQs from Lynn and Becker (2019b) for all three countries are between 60.00 and 62.83, and the QNW IQs are 76.06 (Botswana), 61.95 (Ghana), and 79.59 (South Africa). All of these QNW IQs are within the range of *z*-scores generated from PIRLS and TIMSS data, even though they were derived from independent data.[22]

The SACMEQ is especially useful for comparing scores with others because it shares a number of items with the TIMSS test that is periodically administered to economically developed nations. This permits scores on the SACMEQ to be equated to the scores on TIMSS, and both results are put on the same scale. When Sandefur (2018) did this for the 2007 version of the SACMEQ test, he found that the mean for most SACMEQ nations was at or below the level of a child at the 5th percentile in the UK, a level that is the equivalent of an IQ of 75.3 (i.e., a *z*-score of -1.65 compared to the British population). Indeed, the average child in Malawi, Namibia, and Zambia (the three lowest-scoring nations in the 2007 SACMEQ) scored about 3 standard deviations below the US average on the TIMSS scale (equal to an IQ of 55). This shows that the SACMEQ does not just provide convergent validity regarding the relative rank order of countries within the southern and eastern African region, but it also shows that extremely low national IQs are plausible, despite the critics' incredulity (e.g., Ebbesen, 2020).

Intra-national Variability of IQs

It is important to note, though, that some intranational discrepancies were extremely large; six countries had IQs based on Patel and Sandefur's (2020) data that differed from the QNW + SAS IQs by 10 points or more.[23] When the geographically imputed IQs are included, the number

[22] The QNW + SAS IQs for these countries are 69.45 (Botswana), 60.98 (Ghana), and 69.80 (South Africa). However, note that these are not independent of the PIRLS and TIMSS data because Lynn and Becker (2019b) used the educational achievement data to calculate SAS IQs, which contributed data to the QNW + SAS IQs.

[23] The largest discrepancies were for the Dominican Republic (+ 20.11 IQ points), Yemen (+ 19.34 IQ points), Tunisia (+ 12.39 IQ points), Argentina (+ 11.45 IQ points), Kuwait (+ 10.59 IQ points), and Honduras (− 10.47 IQ points). In this list, positive numbers indicate a higher QNW + SAS score in the Lynn and Becker (2019b) dataset, and negative numbers indicate a higher IQ derived from the Patel and Sandefur (2020) study.

increases to 10 countries with a discrepancy of at least 10 IQ points.[24]

As would be expected, there are intra-country discrepancies between the IQs derived from Gust et al.'s (2022) data and Lynn and Becker's (2019b) IQs, but they are small. The QNW IQs are, on average 0.59 IQ points higher (median = 1.55 IQ points, SD = 7.39 IQ points). QNW + SAS IQs are an average of 1.20 IQ points higher (median = 1.51 IQ points, SD = 6.11 IQ points) in Lynn and Becker's (2019b) data. These averages mask some large discrepancies, though. Fifteen countries (out of 159, 9.4%) had differences of 10 IQ points or more between the IQs derived from the Gust et al. (2022) data and the Lynn and Becker (2019b) NQW + SAS IQs.[25]

Intra-national differences between Lynn and Becker's (2019b) scores and the HLO-derived IQs from the Angrist et al. (2021) data produce similar results. The Lynn and Becker (2019b) dataset produces IQs that are an average of 3.39 IQ points lower for the QNW IQs (SD = 11.6 IQ points, median = 2.29 IQ points lower), 3.18 IQ points lower for the SAS IQs (SD = 3.78, median = 2.13 IQ points lower), 2.41 IQ points lower for QNW + SAS IQs (SD = 6.23 IQ points, median = 2.41 points lower), and 1.90 IQ points lower for the QNW + SAS + GEO IQs (SD = 7.34 IQ points, median = 2.13 IQ points lower). Like the regional and multiregional validations, there are some large discrepancies—even though the average discrepancy is small. Four countries had QNW + SAS IQs in Lynn and Becker's (2019b) dataset that were 10 points or lower than the HLO-derived IQs.[26]

[24] The four countries with geographically imputed IQs in Lynn and Becker's (2019b) dataset that have discrepancies of at least 10 IQ points are Paraguay (+17.26 IQ points), Senegal (−15.76 IQ points), Chad (+13.92 IQ points), and Niger (+10.10 IQ points). In this list, positive numbers indicate a higher QNW + SAS + GEO score in the Lynn and Becker (2019b) dataset, and negative numbers indicate a higher IQ derived from the Patel and Sandefur (2020) study.

[25] The largest discrepancies were for Cambodia (+26.4 IQ points), Venezuela (−23.1 IQ points), Cuba (−20.6 IQ points), Pakistan (+18.4 IQ points), Nicaragua (−15.9 IQ points), Sri Lanka (+15.9 IQ points), Guatemala (−15.4 IQ points), the Dominican Republic (+15.3 IQ points), the Philippines (+14.8 IQ points), Kyrgyzstan (+13.1 IQ points), Argentina (+12.4 IQ points), Haiti (+12.2 IQ points), Morocco (−11.4 IQ points), Mongolia (+10.8 IQ points), and the United Arab Emirates (−10.1 IQ points). In this list, positive numbers indicate a higher QNW + SAS score in the Lynn and Becker (2019b) dataset, and negative numbers indicate a higher IQ derived from the Gust et al. (2022) study. The inclusion of Cuba on this list is due to the use of SERCE 2006 data in the Gust et al. (2022) paper. As I stated earlier in this article, the Cuban data for this test are an outlier and likely fraudulent. This shows that when national IQ discrepancies arise in different datasets, it does not always indicate that Lynn and Becker's (2019b) data are wrong.

[26] In descending order of the magnitude of the discrepancy, these countries were Honduras (22.62 IQ points lower), Botswana (18.52 IQ points lower), South Africa (13.80 IQ points lower), and Egypt (11.65 IQ points lower).

Low National IQs and Intellectual Disability

Some critics still find low sub-Saharan African IQ estimates from Lynn and Becker (2019b) to be unrealistic because low values would imply that many—sometimes a majority—of people in these countries would have an intellectual disability (e.g., Ebbesen, 2020; Sear, 2022). However, this argument betrays a lack of understanding of the nature of the intellectual disability. According to the fifth edition of the *Diagnostic and Statistical Manual for Mental Disorders* (American Psychiatric Association, 2013, p. 37), a person cannot be diagnosed with an intellectual disability without meeting three criteria: (1) a deficit in general mental ability, (2) impairment in functioning compared to one's peers, and (3) onset during childhood. A low IQ score is only relevant for criterion (1), and so a person—or an entire group of people—cannot be judged to have an intellectual disability on the sole basis of a low IQ score. All three criteria must be met because a low IQ can occur for many reasons, such as a language barrier, low motivation, and malingering.

Even if the average IQ for a nation really were extremely low by Western standards, it is not logically possible for a substantial portion of a nation's citizens to have an intellectual disability. This is because the second criterion requires a person's functioning to be impaired compared to that of their peers. By definition, a nation's population of people is a set of peers, and the people within that population cannot logically have an impairment compared to their own functioning. Therefore, low mean national IQ estimates in the Lynn and Becker (2019b) dataset are not evidence—by themselves—that the scores are incorrect simply because they are low enough that a person in a Western country with the same score would be a candidate for a diagnosis of an intellectual disability.

Skills and Functioning in Countries with Low National IQs

Still, it is worth asking whether the low IQ scores in African countries translate into a substantially lower level of functioning than found in high-scoring countries. Again, the international educational testing data are a useful source of data for this because the scores on these tests are anchored to benchmarks that describe the educational competence of examinees. For example, for the fourth grade 2019 TIMSS test, students who meet a "low international benchmark" (a score of 400, which is one standard deviation below the average of 500) can add, subtract, multiply, and divide one-digit and two-digit whole numbers. In economically developed nations, almost all fourth graders can do this: 99% in South Korea and Japan, 96% in England, and 93% in the US. In Morocco, only 43% of fourth graders had this level of math competence. In South Africa, even fewer examinees—37%—reached this basic level of mathematics

competence. Even that percentage is inflated because South Africa administered the TIMSS test to tested fifth graders, whereas almost every other country administered it to fourth graders. Had South Africa tested children in the fourth grade, the percentage would be even lower.[27]

Data from PASEC reveal the same trend. Patel and Sandefur's (2020) data show that several African countries have low percentages of students who can meet the TIMSS low international benchmark for mathematics: Togo (46%), Cameroon (41%), Congo (41%), Benin (35%), Tunisia (35%), Cote d'Ivoire (29%), Chad (17%), and Niger (8%). Among PASEC countries, only Burundi (91%) and Senegal (56%) had more than half of the students above the benchmark.[28] However, all these percentages are inflated compared to fourth-grade data from TIMSS because PASEC countries administer their test in the sixth grade, and these African students have had two more years of schooling to master these concepts.[29] Additionally, in all PASEC countries, only 35.2 to 77.3% of students finish primary school (PASEC, 2015, p. 27). By only testing students enrolled in school, the percentage of children meeting the TIMSS low international benchmarks is inflated in every country. Given these data, it is reasonable to state that the majority of sixth-grade-aged children in at least some sub-Saharan African PASEC countries have not mastered academic concepts that over 90% of fourth graders have mastered in economically developed nations.[30] Such low performance on educational tests is in agreement with the very low IQs that Lynn and Becker (2019b) sometimes find in their dataset and shows just how underdeveloped problem-solving skills are for many people in these countries. Others making international comparisons in academic competence have noticed the same stark discrepancies among students in different countries (e.g., Gust et al., 2022; Kim, 2018; Pritchett & Viarengo, 2021).

The SACMEQ data from southern and eastern Africa also show much lower levels of academic competence in mathematics than in Western countries. For the 2007 administration, "Level 4" mathematics achievement corresponds approximately to the TIMSS low international benchmark. Hungi et al. (2010, pp. 12–23) found that the percentage of students meeting or exceeding the Level 4 mathematics standard was highest in Swaziland (93.0%), followed by Tanzania (89.9%), Kenya (80.2%), Mauritius (78.8%), Seychelles (78.1%), Botswana (75.8%), Namibia (61.3%), Zimbabwe (62.8%), Mozambique (56.5%), Uganda (54.2%), South Africa (51.7%), Lesotho (47.5%), Zambia (27.4%), and Malawi (26.7%). Like the PASEC data, these percentages are inflated by only testing students enrolled in school and by administering the test in the sixth grade instead of the fourth grade. While these percentages are generally higher than those found in the PASEC countries (perhaps due to better educational policies and/or differences in the tests, or the benchmarks), they still show that most of these countries have a substantial portion of early adolescents that fail to reach a level of problem-solving skills that are met or surpassed by over 90% of children in economically developed nations—even though the children in wealthy countries are younger and have not attended school as long.

Exceptionally low performance in sub-Saharan Africa is not limited to children. Sandefur (2018) found that *teachers* in some sub-Saharan African countries had academic achievement levels that were comparable to seventh and eighth graders in high-scoring East Asian and European countries. In a different study of the academic and pedagogical knowledge of elementary school teachers (all of whom had postsecondary training in teaching) in seven sub-Saharan African countries, results indicated that about one-quarter of teachers could not subtract double digits, nearly one-third could not multiply double digits, nearly half could not solve a simple math story problem, and 65% could not solve an algebra problem (Bold et al., 2017, p. 192). These differences are also apparent in non-academic settings, and many adults in economically less developed regions score low on Piagetian tasks or intelligence tests and manifest magical and irrational thinking (Oesterdiekoff, 2012; Rindermann et al., 2014).

Gust et al.'s (2022) goal was to estimate the percentage of adolescents worldwide who had mastered basic skills. This makes their paper especially useful for understanding what extremely low scores (by East Asian or Western standards) mean in the context of learned skills. Gust et al. (2022) defined meeting "basic skills" as obtaining a score at or above the threshold of Level 1 on PISA. This is defined as being able to correctly answer questions that are simple, with clearly defined instructions, and have all necessary information presented in an obvious format. In mathematics,

[27] Testing students one grade higher typical is standard practice for South Africa when administering PIRLS and TIMSS tests.

[28] The Burundi data are clearly an outlier. Patel and Sandefur (2020) reported that 43% of examinees in Burundi met or exceeded the TIMSS low international benchmark in reading, which is typical of PASEC countries (PASEC, 2015, p. 50). The discrepancy between Burundi's math and reading performance originates in the PASEC data and is not an error in Patel and Sandefur's conversion of PASEC scores to TIMSS scores.

[29] Pupil age is another factor to consider in making these comparisons. Repeating a grade is much more common in sub-Saharan Africa than it is in Western countries. However, these older pupils score worse on the PASEC than their classmates who have never repeated a grade (PASEC, 2015, pp. 78–81). Unlike testing students in a higher grade, the inclusion of these older students does not increase the countries' percentages of students who meet the TIMSS low international benchmark.

[30] I only compared mathematics scores here because language differences (e.g., one language being easier to learn to read than another) make comparing reading scores and competency less straightforward than comparing proficiency in mathematics (Gust et al., 2022). Additionally, many children in African learn to read in a non-native language (i.e., Swahili, or a colonial language instead of their local African language), which would be a penalty when comparing reading scores to children in economically developed nations where most children are tested in their native language.

this means students can read a single number from a clearly labeled chart and perform simple arithmetic with whole numbers if given explicit instruction. In science, students at this level can identify simple patterns in data, recognize basic scientific terms, use commonly taught information to recognize the scientific phenomenon, and perform a simple scientific procedure if given clear instructions (OECD, 2019, pp. 109, 117).

Despite the extremely basic level of proficiency needed to master this level of academic competency, Gust et al. (2022, Table 2) estimated that 61.7% of adolescents worldwide enrolled in school and 65.7% of all adolescents were unable to reach PISA level 1 proficiency. In sub-Saharan Africa, 89.3% of students and 94.1% of all adolescents were below this level of proficiency. South Asia does not fare much better: 85.0% of students and 89.2% of all adolescents failed to obtain PISA level 1 skills. The Middle East/North Africa and Latin America/Caribbean regions also have more than half of both groups below this level of proficiency. In fact, in 101 of 159 countries (63.5%) over half of adolescents do not obtain basic skills, and in 36 countries (22.6%), over 90% do not. Conversely, about three-quarters of all adolescents in North America (76.1%) and Europe (71.6%) reach or exceed this level of proficiency. These findings align well with the implied level of academic competence that would inevitably follow from the very low average IQs for many nations in the Lynn and Becker (2019b) dataset.

How to Interpret National IQ Mean Estimates?

As this paper shows, there is a great deal of validity evidence in favor of using and interpreting Lynn and Becker's (2019a, b) national IQ estimates. This evidence consists of (1) external validity data showing that the scores correlate with a wide range of economic, educational, economic, and quality-of-life variables; (2) convergent validity evidence via positive correlations with IQ calculations and educational achievement test scores; and (3) mean scores at the national, regional, and worldwide level that are within the realm of plausibility, as indicated by similar values as mean IQs derived from other sources of data and the alignment of these scores with the observed average cognitive skills of different nations. With all of the validity data reviewed here, it is natural to propose a valid interpretation of national mean IQ estimates.

For some countries, there is probably little controversy over interpreting national IQs as estimates of a country's average intelligence. The USA (QNW + SAS IQ = 97.46) and the UK (QNW + SAS IQ = 99.22) are ideal examples of this scenario. Many of the tests that form the basis of the national IQ estimates were developed in these countries as measures of intelligence, and tests from these nations have a long history of being interpreted as measuring intelligence, with decades of supporting evidence. Given this history, it is reasonable to interpret the UK as having a slightly higher average intelligence than the US.

For some other countries, such an interpretation is questionable at best. For example, most of the studies in sub-Saharan Africa that contributed data to Lynn and Becker's (2019b) national IQ dataset were nonverbal matrix reasoning tests (e.g., Raven's Progressive Matrices). There is strong evidence that in many sub-Saharan African samples, matrix tests do not function the same way as in Western samples (Becker et al., 2022; Dutton et al., 2018; Wicherts et al., 2010a). Thus, I am extremely skeptical of the validity of interpreting national IQs in many of these countries as reflecting the average intelligence of their citizens.

Given the existence of a great deal of validity evidence, it is clear that the national mean IQ estimates measure something; the real dispute is what they measure. A conservative interpretation that I believe can safely be applied to the UW, NW, QNW, SAS, and QNW + SAS IQs (Lynn & Becker, 2019b) is that *the scores measure how well a country's citizens have been trained to solve standardized, abstract problems on tests*. Such an interpretation is firmly grounded in the data and the nature of the tests and makes no assumptions about the underlying construct that the tests measure in any country or population. While this may be unsatisfying to researchers who are interested in the concept of intelligence, some dissatisfaction is a small price to pay for an empirically supported and valid interpretation of important psychometric data.

It is essential to note that this proposed interpretation of the national IQ estimates cannot be supported for the geographically imputed IQs. The weak and/or negative correlations that geographically imputed IQs show with educational achievement data from the same countries (with the exception of Gust et al., 2022, data) is a major deficiency in any valid interpretation of these scores as measures of test performance. Likewise, interpreting geographically imputed IQs as measures of any form of cognitive competence is probably not justified.

Problems

Despite the evidence in favor of Lynn and Becker's (2019b) estimated national IQ means, there are shortcomings of the dataset that scientists should recognize. These shortcomings relate to the compilation of the dataset, the underlying data used to calculate the means, and questionable aspects of the results themselves.

Data Collection Procedures

One of the fundamental reporting standards for a meta-analysis is that the search procedures are adequately described and documented (American Psychological Association Publications & Communications Board Working Group on

Journal Article Reporting Standards, 2008). Therefore, it is a major shortcoming that Lynn and Becker (2019a) did not describe their methodology for searching for IQ scores. Even basic questions—for example, the languages that searches occurred in, search terms the authors used, the time frame that studies occurred in, and whether there were any efforts to access unpublished data sources like dissertations—are not mentioned. In short, people who use Lynn and Becker's (2019b) dataset do not know how comprehensive the underlying data are or whether search procedures might systematically favor some research literature (e.g., articles published in English). This is a criticism that was been made of earlier versions of the national IQ dataset (Wicherts et al., 2010a, c), and it still applies to the most recent version (Lynn & Becker, 2019b).

Likewise, inclusion and exclusion criteria—another basic piece of information that should be reported in a meta-analysis—are not mentioned in Lynn and Becker's (2019a) methodology. Some inclusion criteria can be inferred (e.g., the requirement of an average sample score on an intelligence test), as can some exclusion criteria (e.g., clinical samples were eliminated), but a full list of criteria is not available. As a result, when a dataset surface that is not included in Lynn and Becker's (2019b) work, it is unclear whether this is because (1) it was missed in the search procedures, (2) it did not meet inclusion criteria, or (3) a subjective decision led to it being eliminated. Wicherts et al. (2010a, b, c, d) drew attention to this problem in an earlier version of the national IQ dataset, and it is unclear whether Lynn and his colleagues have taken any steps to remedy this inadequacy.

Indeed, it is not difficult to find samples that could potentially be included in the Lynn and Becker (2019b) data but are missing. In my work on related topics, I have stumbled upon many reports of samples (e.g., Attallah et al., 2014; Bakhiet et al., 2017; Bhatia, 1955; Flores-Mendoza et al., 2018; Gichuhi, 1999; Haile et al., 2016; Irvine, 1964; Lean & Clements, 1981; MacArthur et al., 1964; McFie, 1954; Miezah, 2015; Panza Lombardo, 2016; Ruffieux et al., 2009; Sen et al., 1983; Songy, 2007; van den Briel et al., 2000; Zhao et al., 2019) that could potentially be part of the Lynn and Becker (2019b) dataset. Wicherts et al. (2010a, pp. 138, 139, c, pp. 7–8, d, p. 33) also provided lists of samples from sub-Saharan Africa, many of which do not appear in the Lynn and Becker (2019b) datasets. Additionally, international academic achievement data from before 1995 are missing, even though there are international assessments that predate that year (see, for example, Hanushek & Kimko, 2000). On the other hand, pre-2000 international educational achievements have been criticized for being low quality (Angrist et al., 2021).

One data source that is heavily represented in the Lynn and Becker (2019b) dataset is scores on the Raven's matrix test.[31] This test was designed to measure intelligence with a 3×3 grid of nonverbal stimuli that the user must complete by selecting the option that matches the pattern of images (Raven et al., 1998). Its high factor loading makes it one of the best measures of intelligence (Jensen, 1980), and matrix tests have become a widely popular item format on intelligence test batteries. Matrix tests have found widespread use in cross-cultural testing because only the instructions need to be translated (not the stimuli), they require no verbal or written responses if administered individually, and their simple geometric patterns are not very culturally specific (though, it must be recognized that such shapes are not culturally universal). In the Lynn and Becker (2019b) dataset, matrix tests were used to collect data for 475 of 683 samples (69.5%), of which 474 were Raven's tests.

In the context of cross-national comparisons, the Raven's test has presented problems. Multiple studies have shown that the test does not function the same way in sub-Saharan African countries as it does in Western populations (Becker et al., 2022; Dutton et al., 2018). In contrast to the strong factor loadings in Western nations, the Raven's test is a much weaker indicator of intelligence in sub-Saharan Africa (Wicherts et al., 2010a). Consequentially, the Raven's tests—and other matrix tests—likely underestimate cognitive functioning in sub-Saharan Africa (and possibly in other impoverished nations). The reliance on matrix tests and their poor functioning in some groups is a major problem for the Lynn and Becker (2019b) dataset (Wicherts et al., 2010c), and users should be skeptical of low mean IQ scores based solely on matrix test data.[32] On the other hand, Table 2 shows that IQs that are derived from educational tests are similar to IQs in Lynn and Becker's (2019b) data, which means that the latter dataset should not be dismissed because of its reliance on matrix tests.

Another legitimate criticism is that the quality of the data varies, which is a problem of many meta-analyses. While the variability in data quality probably does not systematically bias the IQ estimates, it is a problem when a country has a small number of samples that contribute to IQ estimates. In Lynn and Becker's (2019b) dataset, the average country only has 5.21 datasets that contribute to UW, NW, or QNW IQs (median = 3, SD = 6.61). For most countries, a single low-quality dataset is enough to reduce accuracy in the final estimate. Weighting by quality—as occurs in the QNW IQ—and including academic achievement scores—such as in the QNW + SAS IQ—likely reduce the severity of this problem but will not eliminate it completely.

Another issue with data quality is that there is legitimate room for debate about what characteristics of samples constitute "high quality." Lynn and Becker (2019b) based their quality

[31] There are three versions of the Raven's: the Colored Progressive Matrices, Progressive Matrices, and Advanced Matrices (listed in ascending order of difficulty).

[32] Countries with a low NWQ+SAS IQ (≤75) based solely on matrix test data are Benin, the Republic of the Congo, Djibouti, Dominica, Eritrea, Ethiopia, The Gambia, Guatemala, Malawi, Mali, Morocco, Namibia, Nepal, Saint Vincent and the Grenadines, Sierra Leone, Somalia, South Sudan, Syria, Tanzania, Yemen, and Zimbabwe.

metric on sample characteristics, test characteristics, and score calculation procedures. While this is helpful information, this is not a comprehensive list of influences on sample quality. Test administration (e.g., whether the administrator is a professional examiner, individual vs. group test administration), test characteristics (e.g., test translated into a new language or adapted to another culture, reading and/or writing requirements for the test), testing location (e.g., school, clinic, home, government office, workplace), and examinee characteristics (e.g., literacy level, number of years of formal schooling) may be important to code and include in future quality metrics.

Another shortcoming of the dataset is the not-uncommon discrepancy between SAS IQs and the IQs derived from intelligence test data (i.e., UW, NW, and QNW IQs). At the individual level, IQ is a strong predictor of school performance in Western countries, and yet large discrepancies between academic achievement data and IQ data occur in Lynn and Becker's (2019b) dataset. This is true whether the academic achievement scores are Lynn and Becker's (2019b) own SAS IQs, or scores derived from other research (see Table 2). Some causes of discrepancies may not be concerning; for example, if school achievement data are not representative because school attendance rates are low, then the IQ data may better reflect a country's problem-solving training. Other discrepancies may be an indication that the IQ data are unrealistic, such as extremely low scores on Raven's tests. These discrepancies should be evaluated on a nation-by-nation basis, and it seems no one has done so yet. Leveraging Lynn and Becker's (2019b) quality ratings may be helpful for this purpose.

Another shortcoming is that the reliability of the underlying scores is often unknown. Like validity, reliability is not a property of the tests themselves. Instead, reliability is a property of a particular set of scores (AERA et al., 2014, Chapter 2), and a test can produce highly reliable scores in one population and highly unreliable scores in another. Most researchers do not report reliability for their own data and instead assume that the reliability values from a previous set of scores will apply to new data—an erroneous practice called reliability induction (Vacha-Haase et al., 2000). It would nonetheless be helpful for score reliability to be incorporated into the quality ratings of the samples in Lynn and Becker's (2019b) dataset.

Finally, the analyses on educational achievement scores that were collected from countries lacking IQ data show that scores derived from neighboring countries are often poor stand-ins for mean national IQ estimates. Given that 52 of 201 (25.9%) countries have a geographically imputed score as their only estimated IQ, this is a problem for a substantial number of nations.

Recommendations

The previous section shows that there is legitimate concern about some of the underlying data that contribute to Lynn and Becker's (2019b) estimates for national mean IQs. Concerns about underlying data—and the consequences thereof—are common in meta-analyses. As previous scholars have written about meta-analysis, "We may not like the ingredients that go into making this sausage, but the [meta-analysis] chef can only work the ingredients provided by the literature" (Thompson & Vacha-Haase, 2000, p. 184). I believe that, given the ingredients in their meta-analyses, many of the estimated mean country IQs in Lynn and Becker's (2019b) dataset approximate the best values that can be derived. However, there is still room to improve this meta-analysis recipe by selecting better ingredients and combining them in better ways. In this section of the article, I propose remediations to the problems I have identified in Lynn and Becker's (2019a, b) work.

Long-term Solutions

Fundamentally, the best thing that scientists can do in response to the shortcomings of the Lynn and Becker (2019b) dataset is to collect data to improve the estimates of national IQs. This may involve identifying pre-existing datasets or published summary statistics that can be included in meta-analyses of national IQ averages or collecting original data in underrepresented countries. Indeed, given the sparse or non-existent IQ data for many nations in the developing world, collecting new IQ data is a great opportunity for cross-cultural psychologists, anthropologists, and researchers in developing nations to make important contributions to the scientific understanding of international differences in problem-solving mastery. Lynn and Becker's (2019a, b) work is practically a guidebook for identifying the greatest needs—in terms of geography and sample characteristics—for cross-national and cross-cultural research on intelligence, cognitive development, and educational achievement. Researchers who broaden the array of tests used to calculate national mean IQs (e.g., Wicherts et al., 2010c) would also be making valuable contributions. In an ideal world, an international organization would create a cross-culturally applicable PISA- or TIMSS-like intelligence test that would permit easy international comparisons of IQ scores and administer this test to many countries.

Additionally, I recommend more research on understanding how tests function in different cultural environments. Score comparisons across populations are only interpretable as reflecting differences in the underlying construct if measurement invariance can be shown. Unfortunately, tests of measurement invariance across countries are lacking—especially when such tests include populations from the developing world. Efforts into testing measurement invariance of intelligence tests in impoverished nations have begun (see Holding et al., 2018; Warne, 2022), but many more studies are required. Until this occurs—and the results indicate a high degree of invariance between populations—it is not justified to say that mean IQ score differences across nations reflect mean differences in intelligence (Hunt & Carlson, 2007). For this reason, I believe

that the most valid interpretation of Lynn and Becker's (2019b) mean national IQ estimates is as a measure of how well populations have been trained to solve the formal problems that appear on tests.

Another long-term solution would be for Lynn and his colleagues (or any other researchers willing to take on the challenge) to bring their work up to the standards of meta-analyses. This would require a systematic search procedure and improved documentation (especially regarding the inclusion and exclusion criteria) and reporting a wider variety of descriptive statistics for each sample—especially standard deviations. Additionally, a list of excluded datasets and the rationale for this decision would improve the transparency of Lynn and his colleagues' work. Perhaps a method of suggesting datasets or articles for consideration would allow outsiders to contribute data and reduce the level of noncoverage for many countries.

Short-term Remediation

Collecting new data and conducting tests of measurement invariance are long-term solutions to improving the quality of the national mean IQ estimates. In the meantime, there are short-term options for researchers who believe that the dataset contains important data that can be used to answer research questions or test hypotheses.

First, a researcher can recalculate any means after eliminating samples that do not meet the researcher's standards for data quality. This is not difficult because Lynn and Becker (2019b) provide a spreadsheet that researchers can download and edit according to their needs. The spreadsheet and the accompanying explanatory book (Lynn & Becker, 2019a) also describe every sample, any unique decisions that the compilers made about the samples, and references to the data's original publication. Thus, all the tools are available to anyone who wants to audit the data and identify samples that meet a higher threshold of quality.

Second, there is the simple option of deleting estimated IQs from countries that have little or no data. Indeed, when a country only has a geographically imputed IQ, I recommend this option. The weak or negative correlations between these geographically imputed IQs and the achievement test data from some of those countries is a major cause of concern and makes it hard to identify any valid interpretation of these IQ scores.[33]

Third, there is the option of Winsorizing low IQs to a minimum that a researcher finds more plausible. Critics of the dataset do not seem to be aware that Lynn and Becker already do this, with 60 sets as the minimum allowable IQ. If users find this value too low (as I do), their spreadsheet (Lynn & Becker, 2019a, b) has an option where users can input a minimum acceptable IQ, and all countries with a lower estimated IQ are automatically reassigned to the new minimum value. All statistics are then automatically recalculated accordingly. For example, with a minimum IQ set at 60, the worldwide mean IQ is calculated at 86.72. Winsorizing the data and setting the minimum value to 75 increases this worldwide mean to 88.30. Finding this new mean was a simple matter of changing the value of one cell in a Microsoft Excel spreadsheet (Lynn & Becker, 2019b).

An Inadequate Option

One final way to cope with the shortcomings of Lynn and Becker's (2019b) dataset is to dismiss or reject it completely. This option is short-sighted and inadequate for two reasons. First, it ignores the validity data that have accumulated regarding the mean IQ scores. Clearly, the national mean IQ estimates measure something of importance.[34] Otherwise, they would not correlate with economic, health, and educational data. I agree that this "something" is not necessarily intelligence—especially in economically developing nations. But automatically rejecting the entire dataset because some of the scores are flawed throws the figurative baby out with the bathwater.

Indeed, one of the laziest foundations for rejecting the estimated mean national IQs is to identify a few poor-quality samples and use their inadequacies to argue that all of Lynn and colleagues' IQs are inadequate. This is a feeble exercise for two reasons. First, none of the authors who have done this (e.g., Dickins et al., 2007; Ebbesen, 2020; Kamin, 2006; Sear, 2022; Wicherts et al., 2010a) have publicly assessed every sample contributing to the national IQ dataset, leaving these critics open to accusations of cherry-picking. Second, comparing older criticisms of this type to the list of samples used in the most recent version of the dataset (Lynn & Becker, 2019b) shows that many of the most criticized samples are no longer included in the IQ calculations. Isolated criticism of samples can quickly become outdated, and if Lynn and his colleagues agree that the samples are too flawed, then dropping them quickly fixes the problem—making the new dataset better than its predecessors.

The second reason that rejecting Lynn and Becker's (2019b) data is inadequate is that doing so inhibits scientific understanding and discovery. The scientific questions that the dataset can help answer are important to researchers in psychology, sociology, economics, education, public health, and many other fields. Denying researchers in so many fields the use of

[33] This is why I have preferred to use the QNW + SAS IQs whenever possible in this article. QNW + SAS IQs are based on the most data and do not include countries with geographically imputed mean IQs.

[34] That is, unless one does not believe that educational performance, life outcomes, health and disease, economic prosperity, and strong civic institutions are important.

the national IQ dataset will make answering these questions more difficult. While the data are imperfect and vary in quality, this is true of many international datasets (Kim, 2018). Imperfect answers to important questions are better than no answers at all.

I recognize that the Lynn et al. estimated mean IQ scores can sometimes produce controversial findings. This is not a sufficient reason to dismiss the dataset or hold it to a higher standard than other international data or other meta-analyses. In fact, the potential for controversy should spur researchers to improve the quality of data so that controversies can be addressed as objectively as possible. Yet, instead of working to improve the quality of national IQ research, the critics seem content to dismiss the data. This is an inadequate response. As I have stated elsewhere:

> It is telling that scholars who demand extra care for controversial intelligence research do not design studies that meet their standards to investigate controversial topics. If their concern for methodological rigor were fully genuine, they would conduct the studies that they demand. This would be the only way to both answer important scientific questions and ensure that those answers are based on trustworthy data. However, the critics never seem interested in collecting the data that they demand from others (Warne, 2020, p. 293).

Moreover, no amount of controversy will change the fact that calculating mean IQs for countries is a scientifically worthwhile endeavor and that there is no objective reason to reject the results of such an effort. National estimated mean IQs are just like any other meta-analysis, and each country's score should be evaluated on the basis of technical adequacy, without regard to scientifically irrelevant criteria. Sometimes the most controversial topics are the most worthwhile to investigate.

Conclusion

Richard Lynn and his colleagues have worked to compile what might be one of the most contentious datasets in the social sciences. Reactions have often been simplistic: either wholesale acceptance or rejection of the data. In this article, I critically evaluated the most recent version of Lynn's national mean IQ estimates (Lynn & Becker, 2019b) and showed that, as a whole, the scores in the dataset correlate with a wide variety of national-level data. This is strong evidence that they measure something important.

However, that does not mean that the scores in Lynn and Becker's (2019b) dataset can be used without reservations. The dataset is very heterogeneous, with data quality, sources, and even the scores themselves being highly variable. Users should evaluate the data and be prepared to make transparent, justifiable decisions about which scores to use, drop, and adjust (e.g., through Winsorizing). Because these decisions have some subjectivity in them, pre-registering these decisions will help increase the transparency and trustworthiness of researchers' results.

In the long term, I hope that this article spurns another round of improvement in the dataset. Meeting the reporting standards for meta-analyses, broadening the range of psychometric tests, improving geographic representation, identifying and coding more aspects of data quality, and understanding how scores function across populations would all improve the value of estimated mean IQ scores even further. With time, I believe that national mean IQ estimates can be a valuable tool for understanding cross-cultural psychology and the nature and magnitude of international differences in a variety of group-level outcomes.

National IQ estimates may not be perfect, but when used thoughtfully and interpreted conservatively, they can be instrumental in testing theories and providing insights that would otherwise be unavailable. I urge social scientists to reject the false dichotomy of accepting mean national IQ scores uncritically or rejecting the entire dataset.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40806-022-00351-y.

Author Contributions The author is solely responsible for the theoretical analysis, identification of secondary data, analysis of data, and writing of the manuscript.

Data Availability Data for new analyses in this article are available in the accompanying Supplemental file.

Declarations

Ethics Approval No ethics approval was necessary because the article does not describe human subjects' research. All data analyzed were publicly available archival data.

Consent to Participate Informed consent to participate was not necessary because the article does not describe human subjects' research. All data analyzed were publicly available archival data.

Consent to Publish All data in this article were aggregate data publicly available from online datasets. Consent to publish is not necessary.

Conflict of Interest The author declares no conflict of interest.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual for Mental Disorders* (5th ed.). American Psychiatric Publishing.

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9), 839–851. https://doi.org/10.1037/0003-066x.63.9.839

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

Attallah, S. E. F., Ahmed, K. Y., & Meisenberg, G. (2014). Factor structure of Wechsler Intelligence Scale for Children-Third Edition (WISC-III) among gifted students in Sudan. *Mankind Quarterly, 55*, 147–170. https://doi.org/10.46469/mq.2014.55.1.11

Bakhiet, S. F., Albursan, I. S., Al Qudah, M. F., Abduljabbar, A. S., Aljomaa, S. S., Toto, H. S. A., & Lynn, R. (2017). Sex differences on the WISC-III among children in Sudan and the United States. *Journal of Biosocial Science, 49*(6), 792–797. https://doi.org/10.1017/s0021932016000432

Barnes, J. C., Beaver, K. M., & Boutwell, B. B. (2013). Average county-level IQ predicts county-level disadvantage and several county-level mortality risk rates. *Intelligence, 41*(1), 59–66. https://doi.org/10.1016/j.intell.2012.10.007

Bauer, P. J. (2020). A call for greater sensitivity in the wake of a publication controversy. *Psychological Science, 31*(7), 767–769. https://doi.org/10.1177/0956797620941482

Beaver, K. M., Schwartz, J. A., Connolly, E. J., Said Al-Ghamdi, M., Kobeisy, A. N., Barnes, J. C., & Boutwell, B. B. (2016). Intelligence and early life mortality: Findings from a longitudinal sample of youth. *Death Studies, 40*(5), 298–304. https://doi.org/10.1080/07481187.2015.1137994

Becker, D., Meisenberg, G., Dutton, E., Bakhiet, S. F., Humad, O. A. M., Abdoulaye, H. A., & Ahmed, S. A. E. S. (2022). Factor structure in Raven's Progressive Matrices Plus in sub-Saharan Africa – Benin and Djibouti. *Journal of Psychology in Africa, 32*(2), 103–114. https://doi.org/10.1080/14330237.2022.2028080

Becker, D., & Rindermann, H. (2016). The relationship between cross-national genetic distances and IQ-differences. *Personality and Individual Differences, 98*, 300–310. https://doi.org/10.1016/j.paid.2016.03.050

Belasen, A., & Hafer, R. W. (2013). IQ and alcohol consumption: International data. *Intelligence, 41*, 615–621. https://doi.org/10.1016/j.intell.2013.07.019

Bernstein, B. O., Lubinski, D., & Benbow, C. P. (2019). Psychological constellations assessed at age 13 predict distinct forms of eminence 35 years later. *Psychological Science, 30*, 444–454. https://doi.org/10.1177/0956797618822524

Bhatia, C. M. (1955). *Performance tests of intelligence under Indian conditions*. Oxford University Press.

Bold, T., Filmer, D., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J., & Wane, W. (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa. *Journal of Economic Perspectives, 31*(4), 185–204. https://doi.org/10.1257/jep.31.4.185

Chinapah, V., H'ddigui, E. M., Kanjee, A., Falayajo, W., Fomba, C. O., Hamissou, O. Rafalimanana, A., & Byomugisha, A. (2000). *With Africa for Africa: Towards quality education for all. 1999 MLA Project*. Human Sciences Research Council. http://files.eric.ed.gov/fulltext/ED444931.pdf

Clark, G. (2007). *A farewell to alms: A brief economic history of the world*. Princeton University Press.

Deary, I. J., Pattie, A., & Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian Birth Cohort of 1921. *Psychological Science, 24*(12), 2361–2368. https://doi.org/10.1177/0956797613486487

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*, 130–147. https://doi.org/10.1037/0022-3514.86.1.130

Dickins, T. E., Sear, R., & Wells, A. J. (2007). Mind the gap(s) in theory, method and data: Re-examining Kanazawa (2006). *British Journal of Health Psychology, 12*(2), 167–178. https://doi.org/10.1348/135910707X174339

Dutton, E., Becker, D., Osman, H. A., Bakhiet, S. F., Essa, Y. A. S., Ali, H. A. A., Alqafari, S. M., Hamdi, A. H. M. N., & Alfaleh, A. S. H. (2018). The Raven's test performance of South Sudanese samples: A validation of criticisms of the utility of Raven's among Sub-Saharan Africans. *Personality and Individual Differences, 128*, 122–126. https://doi.org/10.1016/j.paid.2018.02.018

Ebbesen, C. L. (2020). Flawed estimates of cognitive ability in Clark et al. *Psychological Science*, 2020. *PsyArXiv*. https://psyarxiv.com/tzr8c

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

Floyd, R. G., Reynolds, M. R., Farmer, R. L., & Kranzler, J. H. (2013). Are the general factors from different child and adolescent intelligence tests the same? Results from a five-sample, six-test analysis. *School Psychology Review, 42*, 383–401. https://doi.org/10.1080/02796015.2013.12087461

Gale, C. R., Batty, G. D., Tynelius, P., Deary, I. J., & Rasmussen, F. (2010). Intelligence in early adulthood and subsequent hospitalization for mental disorders. *Epidemiology, 21*(1), 70–77. http://www.jstor.org/stable/25662808

Gichuhi, J. (1999). *An examination of the Wechsler Intelligence Scale for Children-III for predicting performance on a national examination of grade eight children in public primary schools* (Unpublished doctoral dissertation). Biola University, La Mirada, CA.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Gust, S., Hanushek, E. A., & Woessmann, L. (2022). *Global universal basic skills: Current deficits and implications for world development* (Working paper No. 30566). National Bureau of Economic Research. https://www.nber.org/papers/w30566

Haier, R. J. (2017). The neuroscience of intelligence. *Cambridge University Press*. https://doi.org/10.1017/9781316105771

Haile, D., Gashaw, K., Nigatu, D., & Demelash, H. (2016). Cognitive function and associated factors among school age children in Goba Town, South-East Ethiopia. *Cognitive Development, 40*, 144–151. https://doi.org/10.1016/j.cogdev.2016.09.002

Hanushek, E. A. (2016). Will more higher education improve economic growth? *Oxford Review of Economic Policy, 32*, 538–552. https://doi.org/10.1093/oxrep/grw025

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review, 90*(5), 1184–1208. https://doi.org/10.1257/aer.90.5.1184

Hart, C. L., Taylor, M. D., Davey Smith, G., Whalley, L. J., Starr, J. M., Hole, D. J., Wilson, V., & Deary, I. J. (2003). Childhood IQ, social class, deprivation, and their relationships with mortality and morbidity risk in later life: Prospective observational study linking the Scottish Mental Survey 1932 and the Midspan Studies. *Psychosomatic Medicine, 65*(5), 877–883. https://doi.org/10.1097/01.psy.0000088584.82822.86

Hart, C. L., Taylor, M. D., Smith, G. D., Whalley, L. J., Starr, J. M., Hole, D. J., Wilson, V., & Deary, I. J. (2004). Childhood IQ and cardiovascular disease in adulthood: Prospective observational study linking the Scottish Mental Survey 1932 and the Midspan studies. *Social Science & Medicine, 59*(10), 2131–2138. https://doi.org/10.1016/j.socscimed.2004.03.016

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083–1101. https://doi.org/10.1037/0003-066X.47.9.1083

Holahan, C. K., & Sears, R. R. (1995). *The gifted group in later maturity*. University of Stanford Press.

Holding, P., Anum, A., van de Vijver, F. J., Vokhiwa, M., Bugase, N., Hossen, T., Makasi, C., Baiden, F., Kimbute, O., & Bangre, O. (2018). Can we measure cognitive constructs consistently within and across cultures? Evidence from a test battery in Bangladesh, Ghana, and Tanzania. *Applied Neuropsychology: Child, 7*, 1–13. https://doi.org/10.1080/21622965.2016.1206823

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Cappelle, F., Paviot, L., & Vellien, J. (2010). SACMEQ III project results: Pupil achievement levels in reading and mathematics. SACMEQ. http://www.sacmeq.org/sites/default/files/sacmeq/reports/sacmeq-iii/working-documents/wd01_sacmeq_iii_results_pupil_achievement.pdf

Hunt, E., & Carlson, J. (2007). Considerations relating to the study of group differences in intelligence. *Perspectives on Psychological Science, 2*(2), 194–213. https://doi.org/10.1111/j.1745-6916.2007.00037.x

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

Irvine, S. H. (1964). *A psychological study of selection problems at the end of primary schooling in Southern Rhodesia* (Unpublished doctoral dissertation). University of London.

Jensen, A. R. (1969). Reducing the heredity-environment uncertainty: A reply. *Harvard Educational Review, 39*, 449–483. https://doi.org/10.17763/haer.39.3.4158240700761019

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

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Johnson, W., Bouchard Jr, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence, 32*, 95–107. https://doi.org/10.1016/S0160-2896(03)00062-X

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. Jr. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence, 36*, 81–95. https://doi.org/10.1016/j.intell.2007.06.001

Jones, G. (2016). *Hive mind: How your nation's IQ matters so much more than your own*. Stanford University 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

Kanazawa, S. (2006). Mind the gap in intelligence: Re-examining the relationship between inequality and health. *British Journal of Health Psychology, 11*(4), 623–642. https://doi.org/10.1348/135910705X69842

Keith, T. Z., Kranzler, J. H., & Flanagan, D. P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock-Johnson Tests of Cognitive Ability (3rd edn). *School Psychology Review, 30*, 89–119. https://doi.org/10.1080/02796015.2001.12086102

Kell, H. J., Lubinski, D., & Benbow, C. P. (2013). Who rises to the top? Early Indicators. *Psychological Science, 24*(5), 648–659. https://doi.org/10.1177/0956797612457784

Kim, J. Y. (2018). The human capital gap: Getting governments to invest in people. *Foreign Affairs, 97*(4), 92–101. http://www.jstor.org/stable/44822216

Lean, G., & Clements, M. A. (1981). Spatial ability, visual imagery, and mathematical performance. *Educational Studies in Mathematics, 12*(3), 267–299. https://doi.org/10.1007/bf00311060

Lim, S. S., Updike, R. L., Kaldjian, A. S., Barber, R. M., Cowling, K., York, H., Friedman, J., Xu, R., Whisnant, J. L., Taylor, H. J., Leever, A. T., Roman, Y., Bryant, M. F., Dieleman, J., Gakidou, E., & Murray, C. J. L. (2018). Measuring human capital: a systematic analysis of 195 countries and territories, 1990–2016. *The Lancet, 392*(10154), 1217–1234. https://doi.org/10.1016/S0140-6736(18)31941-X

Lubinski, D. (2009). Exceptional cognitive ability: The phenotype. *Behavior Genetics, 39*(4), 350–358. https://doi.org/10.1007/s10519-009-9273-0

Lubinski, D., Benbow, C. P., & Kell, H. J. (2014). Life paths and accomplishments of mathematically precocious males and females four decades later. *Psychological Science, 25*(12), 2217–2232. https://doi.org/10.1177/0956797614551371

Lynn, R. (2015). *Race differences in intelligence: An evolutionary analysis* (2nd ed.). Washington Summit Publishers.

Lynn, R., & Becker, D. (2019a). *The intelligence of nations*. Ulster Institute for Social Research.

Lynn, R., & Becker, D. (2019b). *The NIQ-Dataset version 1.3.3*. https://viewoniq.org/

Lynn, R., & Meisenberg, G. (2010). National IQs calculated and validated for 108 nations. *Intelligence, 38*, 353–360. https://doi.org/10.1016/j.intell.2010.04.007

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

Lynn, R., & Vanhanen, T. (2006). *IQ and global inequality*. Washington Summit Publishers.

Lynn, R., & Vanhanen, T. (2012). *Intelligence: A unifying construct for the social sciences*. Ulster Institute for Social Research

MacArthur, R. S., Irvine, S. H., & Brimble, A. R. (1964). *The Northern Rhodesia mental ability survey 1963*. Rhodes-Livingstone Institute.

Makel, M. C., Kell, H. J., Lubinski, D., Putallaz, M., & Benbow, C. P. (2016). When lightning strikes twice: Profoundly gifted, profoundly accomplished. *Psychological Science, 27*(7), 1004–1018. https://doi.org/10.1177/0956797616644735

McFie, J. (1954). African performance on an intelligence test. *Uganda Journal, 18*, 34–43.

Miezah, D. (2015). *Validation of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) in the Ghanian population*. University of Ghana.

Murray, C. (2002). IQ and income inequality in a sample of sibling pairs from advantaged family backgrounds. *American Economic Review, 92*(2), 339–343. https://doi.org/10.1257/000282802320191570

Nyborg, H. (2012). The decay of Western civilization: Double relaxed Darwinian selection. *Personality and Individual Differences, 53*(2), 118–125. https://doi.org/10.1016/j.paid.2011.02.031

Oesterdiekhoff, G. W. (2012). Was pre-modern man a child? The quintessence of the psychometric and developmental approaches. *Intelligence, 40*(5), 470–478. https://doi.org/10.1016/j.intell.2012.05.005

Oficina Regional de Educación para América Latina y el Caribe/UNESCO. (2001). *Primer studio internactional comparative sobre lenguaje, matématica y factores asociados, para alumnus del tercer y cuarto grado de la educación básica: Informe técnico*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000149268

Organization for Economic Cooperation and Development. (2016). Skills matter: Further results from the Survey of Adult Skills. *OECD Publishing*. https://doi.org/10.1787/9789264258051-en

Organization for Economic Cooperation and Development. (2019). *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

Panza Lombardo, M. A. (2016). Transferencia y permanencia del entrenamiento dual n-back en la inteligencia fluida y en la memoria de trabajo verbal [Transfer and permanence of dual n-back training on fluid intelligence and verbal working memory]. *Psicodebate, 16*, 49–82.

Patel, D., & Sandefur, J. (2020). *A Rosetta Stone for human capital*. Center for Global Development. https://www.cgdev.org/publication/rosetta-stone-human-capital

Piffer, D. (2013). Correlation of the COMT Val158Met polymorphism with latitude and a hunter-gather lifestyle suggests culture–gene

coevolution and selective pressure on cognition genes due to climate. *Anthropological Science, 121*, 161–171. https://doi.org/10.1537/ase.130731

Pokropek, A., Marks, G. N., & Borgonovi, F. (2022). How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities? *Journal of Educational Psychology, 114*(5), 1121–1135. https://doi.org/10.1037/edu0000687

Pritchett, L., & Viarengo, M. (2021). *Learning outcomes in developing countries: Four hard lessons from PISA-D*. RISE Working Paper No. 21/069. https://doi.org/10.35489/BSG-RISE-WP_2021/069

Programme d'Analyse des Systèmes Éducatifs de la Confemen. (2015). *PASEC2014 education system performance in francophone sub-Saharan Africa: Competencies and learning factors in primary education*. Programme d'Analyse des Systèmes Éducatifs de la Confemen. https://www-pasec-confemen-org.translate.goog/wp-content/uploads/2015/12/Rapport_Pasec2014_GB_webv2.pdf

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales, Section 1: General overview*. Pearson.

Richardson, K. (2002). What IQ tests test. *Theory & Psychology, 12*, 283–314. https://doi.org/10.1177/0959354302012003012

Rindermann, H. (2018a). Cognitive capitalism: Human capital and the wellbeing of nations. *Cambridge University Press*. https://doi.org/10.1017/9781107279339

Rindermann, H. (2018b). *Appendix*. https://www.tu-chemnitz.de/hsw/psychologie/professuren/entwpsy/team/rindermann/pdfs/RindermannCogCapAppendix.pdf

Rindermann, H., Becker, D., & Coyle, T. R. (2017). Survey of expert opinion on intelligence: The FLynn effect and the future of intelligence. *Personality and Individual Differences, 106*, 242–247. https://doi.org/10.1016/j.paid.2016.10.061

Rindermann, H., & Carl, N. (2020). The Good Country Index, cognitive ability and culture. *Comparative Sociology, 19*(1), 39–68. https://doi.org/10.1163/15691330-12341521

Rindermann, H., Falkenhayn, L., & Baumeister, A. E. E. (2014). Cognitive ability and epistemic rationality: A study in Nigeria and Germany. *Intelligence, 47*, 23–33. https://doi.org/10.1016/j.intell.2014.08.006

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A Meta-Analysis. *Psychological Science, 29*(8), 1358–1369. https://doi.org/10.1177/0956797618774253

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Ruffieux, N., Njamnshi, A. K., Mayer, E., Sztajzel, R., Eta, S. C., Doh, R. F., Kengne, A.-M., Ngamaleu, R. N., Chanal, J., & Verdon, V. (2009). Neuropsychology in Cameroon: First normative data for cognitive tests among school-aged children. *Child Neuropsychology, 16*(1), 1–19. https://doi.org/10.1080/09297040902802932

Sandefur, J. (2018). Internationally comparable mathematics scores for fourteen African countries. *Economics of Education Review, 62*, 267–286. https://doi.org/10.1016/j.econedurev.2017.12.003

Schleicher, A. (2019). PISA 2018: Insights and interpretations. Organisation for Economic Co-operation and Development. https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

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

Sen, A., Jensen, A. R., Sen, A. K., & Arora, I. (1983). Correlation between reaction time and intelligence in psychometrically similar groups in America and India. *Applied Research in Mental Retardation, 4*(2), 139–152. https://doi.org/10.1016/0270-3092(83)90006-1

Serpell, R., & Jere-Folotiya, J. (2008). Developmental assessment, cultural context, gender, and schooling in Zambia. *International Journal of Psychology, 43*(2), 88–96. https://doi.org/10.1080/00207590701859184

Sörberg, A., Allebeck, P., & Hemmingsson, T. (2014). IQ and somatic health in late adolescence. *Intelligence, 44*, 155–162. https://doi.org/10.1016/j.intell.2014.04.002

Songy, D. G. (2007). Predicting success in academic achievement of major seminarians in Papua New Guinea: A comparison of cognitive test results and grade point averages. *Contemporary PNG Studies, 7*, 59–71.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. The Macmillan Company.

Starr, J. M., Deary, I. J., Lemmon, H., & Whalley, L. J. (2000). Mental ability age 11 years and health status age 77 years. *Age and Ageing, 29*, 523–528. https://doi.org/10.1093/ageing/29.6.523

Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than *g*: An extension of Kyllonen's analyses. *The Journal of General Psychology, 123*, 193–205. https://doi.org/10.1080/00221309.1996.9921272

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

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

Terman, L. M., & Oden, M. H. (1947). *Genetic studies of genius:* (Vol. IV). Twenty-five years' follow-up of a superior group. Stanford University Press.

Terman, L. M., & Oden, M. H. (1959). *Genetic studies of genius:* (Vol. V). Thirty-five years' follow-up of the superior child. Stanford University Press.

Terracciano, A., Abdel-Khalek, A. M., Ádám, N., Adamovová, L., Ahn, C. K., Ahn, H. N., Alansari, B. M., Alcalay, L., Allik, J., Angleitner, A., Avia, M. D., Ayearst, L. E., Barbaranelli, C., Beer, A., Borg-Cunen, M. A., Bratko, D., Brunner-Sciarra, M., Budzinski, L., Camart, N., & McCrae, R. R. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*(5745), 96–100. https://doi.org/10.1126/science.1117199

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174–195. https://doi.org/10.1177/00131644000602002

United Nations Office on Drugs and Crime. (2022). *Intentional homicide*. https://dataunodc.un.org/sites/dataunodc.un.org/files/data_cts_intentional_homicide.xlsx

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*(4), 509–522. https://doi.org/10.1177/00131640021970682

van den Briel, T., West, C. E., Bleichrodt, N., van de Vijver, F. J. R., Ategbo, E. A., & Hautvast, J. G. A. J. (2000). Improved iodine status is associated with improved mental performance of schoolchildren in Benin. *The American Journal of Clinical Nutrition, 72*(5), 1179–1185. https://doi.org/10.1093/ajcn/72.5.1179

Wai, J. (2014). Experts are born, then made: Combining prospective and retrospective longitudinal data shows that cognitive ability matters. *Intelligence, 45*, 74–80. https://doi.org/10.1016/j.intell.2013.08.009

Warne, R. T. (2020). In the know: Debunking 35 myths about human intelligence. *Cambridge University Press*. https://doi.org/10.1017/9781108593298

Warne, R. T. (2021). *Statistics for the social sciences: A general linear model approach* (2nd ed.). Cambridge University Press.

Warne, R. T. (2022). *Tests of measurement invariance of three Wechsler intelligence tests in economically developing nations in south Asia and sub-Saharan Africa*. Manuscript submitted for publication.

Warne, R. T., & Burningham, C. (2019). Spearman's *g* found in 31 non-Western nations: Strong evidence that *g* is a universal phenomenon. *Psychological Bulletin, 145*(3), 237–272. https://doi.org/10.1037/bul0000184

Warne, R. T., & Burton, J. Z. (2020). Beliefs about human intelligence in a sample of teachers and nonteachers. *Journal for the Education of the Gifted, 43*(2), 143–166. https://doi.org/10.1177/0162353220912010

Warne, R. T., Yoon, M., & Price, C. J. (2014). Exploring the various interpretations of "test bias." *Cultural Diversity and Ethnic Minority Psychology, 20*, 570–582. https://doi.org/10.1037/a0036503

Wolf, T. H. (1973). *Alfred Binet*. The University of Chicago Press.

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & van der Maas, H. L. J. (2010a). 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., Carlson, J. S., & van der Maas, H. L. J. (2010b). Another failure to replicate Lynn's estimate of the average IQ of sub-Saharan Africans. *Learning and Individual Differences, 20*, 155–157. https://doi.org/10.1016/j.lindif.2010.03.010

Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. J. (2010c). A systematic literature review of the average IQ of sub-Saharan Africans. *Intelligence, 38*(1), 1–20. https://doi.org/10.1016/j.intell.2009.05.002

Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. J. (2010d). The dangers of unsystematic selection methods and the representativeness of 46 samples of African test-takers. *Intelligence, 38*(1), 30–37. https://doi.org/10.1016/j.intell.2009.11.003

Williams, R. T., Polanin, J. R., & Pigott, T. D. (2017). Meta-analysis and reproducibility. In *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research.* (pp. 255–270). American Psychological Association. https://doi.org/10.1037/0000033-016

Woodberry, K. A., Giuliano, A. J., & Seidman, L. J. (2008). Premorbid IQ in schizophrenia: A meta-analytic review. *The American Journal of Psychiatry, 165*(5), 579–587. https://doi.org/10.1176/appi.ajp.2008.07081242

Zaboski, B. A., II., Kranzler, J. H., & Gage, N. A. (2018). Meta-analysis of the relationship between academic achievement and broad abilities of the Cattell-Horn-Carroll theory. *Journal of School Psychology, 71*, 42–56. https://doi.org/10.1016/j.jsp.2018.10.001

Zagorsky, J. L. (2007). Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. *Intelligence, 35*, 489–501. https://doi.org/10.1016/j.intell.2007.02.003

Zhao, Q., Wang, X., & Rozelle, S. (2019). Better cognition, better school performance? Evidence from primary schools in China. *China Economic Review, 55*, 199–217. https://doi.org/10.1016/j.chieco.2019.04.005

Zisman, C., & Ganzach, Y. (2022). The claim that personality is more important than intelligence in predicting important life outcomes has been greatly exaggerated. *Intelligence, 92*, Article 101631. https://doi.org/10.1016/j.intell.2022.101631