

Substituting National-IQ and Harmonized-Learning Indicators: An Output-Specific Replication and Sensitivity Audit

Georgios Kiminos

giorgos1973@yahoo.gr

Independent Researcher

Research Article

Keywords: national IQ, learning outcomes, indicator substitution, dependent correlations, rank sensitivity, reproducibility

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10491444/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Country-level cognitive-score datasets are sometimes substituted despite differing sources, assessed populations, periods, and aggregation. We conduct an output-specific replication and sensitivity audit of a direct quality- and recorded-sample-size-weighted national-IQ compilation (QNW), an author-constructed harmonized learning-outcome aggregate (HLO), and the Gust–Hanushek–Woessmann mathematics-and-science achievement indicator (GHW). On the exact 66-economy ranking sample, QNW and HLO correlate highly (Pearson $r=.795$ and Spearman $\rho=.791$). Mean country-level absolute rank displacement divided by $(N-1)$ was .146. Under the analysis-plan empirical-decile rule, each tail contained seven economies. QNW and HLO shared four in the upper tail (Jaccard=.400) and two in the lower tail (Jaccard=.167). On the 65-economy gross domestic product (GDP) sample, QNW and HLO correlate .616 and .690 with log 2017 GDP per capita. Their planned dependent-correlation contrast is $-.074$ (Williams $t(62)=-1.244$, $p=.218$), and the Zou 95% confidence interval (CI) is $[-.207, .045]$. Direct-QNW fixed-sample variants and every leave-one-economy comparison retain negative point contrasts. Coverage-changing indicator-construction contrasts were $-.034$ for balanced HLO, $-.005$ for all-available HLO, and $+.009$ for both school-assessment-augmented constructions. The United Nations (UN) frame contrast was $-.076$ and did not attenuate the primary estimate. Only 65 of 217 economies in the frozen World Bank frame enter the GDP analysis. Population weighting applies to those selected economies and has a Kish weight-only index of 18.9. The audit quantifies output differences but does not establish interchangeability, equivalence, a pure measurement-family effect, or causal relations.

1 Introduction

Country-level score series are used in comparative research for several distinct outputs: describing relative country positions, assigning countries to an extreme group, or estimating an association with a country-level outcome. These uses are often discussed as though a high correlation between two indicators settles whether one can stand in for the other. Correlation measures covariation, not agreement or interchangeability (Bland and Altman 1986; Lin 1989). It neither supplies a tolerance for country-level changes nor specifies the conclusion that must remain stable. Ranking work likewise distinguishes an ordering from the uncertainty or authority attached to it (Goldstein and Spiegelhalter 1996; Hall and Miller 2009). Rank correlations also require explicit treatment of ties (Kendall 1945). Classification overlap is distinct from score correlation. Chance-corrected categorical agreement is another possible summary (Cohen 1960).

This article therefore uses the term *indicator substitution* rather than the broader term *interchangeability*. It asks how specified outputs change when a direct national-IQ compilation is replaced by a specified harmonized learning measure. The outputs are (a) score association, (b) ranks of published country point scores, (c) empirical-tail membership, and (d) the bivariate association with GDP per capita. A nonsignificant difference is not evidence of equivalence. Equivalence testing requires explicit, defensible bounds that should preferably be prespecified (Schuirmann 1987; Lakens 2017). Because no decision loss, smallest effect size of interest, or equivalence margin was specified here, the audit reports quantities and uncertainty without a pass/fail claim that the indicators support the same conclusions.

The primary comparison is between QNW and HLO. QNW is an economy-level compilation assembled from source studies and series metadata, whereas HLO is an aggregate of assessed-student performance harmonized across international and regional assessments. The article adds GHW, the mean mathematics-and-science achievement score reported by Gust et al. (2024). GHW is useful because its linking construction is independent of the focal HLO construction. It is not source-independent in a strong sense: both learning series draw on parts of the international and regional assessment ecosystem. The comparisons consequently describe agreement among published indicators. They do not isolate an IQ-versus-learning effect, establish a common latent trait, or identify a true national ordering.

The two focal research questions (RQs) are explicit. RQ1 (ranking analysis) asks: Within the exact 66-economy intersection of direct QNW and primary HLO, how do score association, ranks based on published country point estimates, and planned empirical-tail membership differ when QNW is replaced by HLO? RQ2 (GDP-association analysis) asks: Within the exact 65-economy complete-case intersection, holding the outcome year, transformation, country set, and equal-economy weighting fixed, how does replacing QNW with HLO change the association with log 2017 GDP per capita? The GHW comparisons are a descriptive extension and are not part of either planned focal contrast.

The contribution is empirical and reporting-oriented rather than a newly validated comparison method. The paper separates source-specific numerical reproductions from new exact-sample sensitivity analyses. It traces substitution through scores, ranks, empirical-tail membership, and shared-outcome associations, and it distinguishes changes in construction, coverage, weighting, and regional composition. The reporting implications draw on dependent-correlation inference (Steiger 1980; Hittner et al. 2003; Zou 2007), specification-sensitivity principles (Steenen et al. 2016; Simonsohn et al. 2020), and reproducibility practice (Stodden et al. 2018; Hardwicke et al. 2018). They are not presented as a new validated framework.

2 Prior claims and audit scope

Direct predecessors used country-level IQ–achievement correlations to make a broader construct argument. Rindermann (2007) reported correlations of .85–.86 between student-assessment composites and national IQ-test compilations and a dominant first factor, interpreting the convergence as evidence for a broad national cognitive-ability factor. Lynn et al. (2007) reported a correlation of .92 between national-IQ estimates and eighth-grade mathematics/science achievement in 67 countries and interpreted it as validation and evidence that the measures represented the same construct. Critical responses identified limits that a large aggregate correlation cannot resolve. Bosker (2007) emphasized that aggregation can inflate country-level associations and that a country-level factor requires a clear multilevel interpretation. Hunt and Wittmann (2008) judged weaknesses in the Lynn–Vanhanen inputs to reduce construct validity and found contemporary student-assessment data more predictive of national prosperity, while cautioning against a simple causal account. Using explicit search and inclusion rules, Wicherts et al. (2010) obtained materially different estimates for sub-Saharan African samples and highlighted unresolved measurement-bias questions. These exchanges establish the gap addressed here: score correlation is one output, not evidence that published ranks, empirical-tail classifications, or associations with an external outcome are preserved.

Numerical reproduction is not substantive endorsement: recovering a statistic under an identified source and specification neither validates the indicator nor makes a new indicator pair equivalent to the published pair. Table 1 separates these tasks for three relevant sources.

Warne (2023) reported Pearson correlations between HLO-derived IQ and variants of the national-IQ (NIQ) dataset. In this terminology, SAS is the School Assessment Score and GEO denotes geographic completion. Using the published coefficients and pairwise sample sizes in the Angrist et al. HLO row of Warne's (2023) Table 1 as the reference, calculations from the publisher supplement's National scores worksheet reproduce the SAS, QNW + SAS, and geographic-only coefficients to three decimals. Two coefficients differ slightly, and several pairwise sample sizes are larger by two or three observations. Undocumented historical exclusions prevent exact reconciliation.

Henss (2024) defines NIQB as Becker's QNW + SAS+GEO, not direct QNW. Its Table 2 uses the Warne archive's IQ-transformed HLO and GHW variables and excludes Cambodia's NIQB value and the learning values for Cuba and Pakistan. Applying those rules reproduces all three displayed coefficients to two decimals. They are not direct-QNW results.

Parra and Kirkegaard (2025) provide public final-score archives for another construction. The files permit audits of rows, identifiers, precision, and sample overlap, but not their GDP correlation because the required four-source 2018–2022 average was absent from the public repository files available on the audit date. Because the series reuses cognate IQ and learning sources, correlations with QNW, HLO, or GHW diagnose source overlap rather than independent confirmation.

Table 1
Prior-claim audit: source-specific specifications and reproducibility status

Source and published claim	Exact indicator pair and specification	Published r	Published N	Reanalysis result	Reproduction status
Warne (2023), Table 1	HLO-derived IQ–QNW, Warne publisher supplement (National scores), pairwise Pearson	.811	112	.809, N = 115	Approximate reproduction (worksheet yields a larger N)
Warne (2023), Table 1	HLO-derived IQ–SAS, Warne publisher supplement (National scores), pairwise Pearson	.966	98	.966, N = 100	Approximate reproduction (worksheet yields a larger N)
Warne (2023), Table 1	HLO-derived IQ–QNW + SAS, Warne publisher supplement (National scores), pairwise Pearson	.871	130	.871, N = 133	Approximate reproduction (Warne's accompanying prose reports .896)
Warne (2023), Table 1	HLO-derived IQ–QNW + SAS+GEO, Warne publisher supplement (National scores), pairwise Pearson	.829	158	.830, N = 161	Approximate reproduction (worksheet yields a larger N)
Warne (2023), Table 1	HLO-derived IQ–GEO only, Warne publisher supplement (National scores), pairwise Pearson	.140	28	.140, N = 28	Numerical reproduction to three decimals
Henss (2024), Table 2	NIQB (QNW + SAS+GEO)–HLO IQ, with KHM omitted from NIQB and CUB/PAK omitted from HLO IQ	.84	not reported	.841196, N = 158	Exact numerical reproduction to displayed precision
Henss (2024), Table 2	NIQB (QNW + SAS+GEO)–GHW IQ, with KHM omitted from NIQB and CUB/PAK omitted from GHW IQ	.89	not reported	.892595, N = 153	Exact numerical reproduction to displayed precision
Henss (2024), Table 2	HLO IQ–GHW IQ, with CUB/PAK omitted from both learning variables	.93	not reported	.928059, N = 148	Exact numerical reproduction to displayed precision

Note. All Warne coefficients are taken from the Angrist et al. (2021) HLO row in Table 1 of Warne (2023). The corresponding pairwise sample sizes appear in the upper triangle. The Henss rows use the archived variables identified in the table, not the direct-QNW, primary-HLO, or GHW variables on the Programme for International Student Assessment (PISA) scale used below. Country codes in Table 1 are KHM (Cambodia), CUB (Cuba), and PAK (Pakistan). SE denotes standard error. These entries concern source-specific numerical reproduction, not indicator equivalence.

Source and published claim	Exact indicator pair and specification	Published r	Published N	Reanalysis result	Reproduction status
Parra and Kirkegaard (2025)	Public final-score archive with 197 records and a mean adjusted SE of 2.58	–	197	197 records in the score table, 198 unique country-code rows in the companion table, and a mean SE of 2.556 (2.56 at two decimals)	Row/identifier discrepancy. The precision summary does not reproduce 2.58 at two decimals.
Parra and Kirkegaard (2025)	National-IQ–GDP correlation using four-source 2018–2022 GDP average	.82	not reported	Not computed	Required GDP source series are absent from the public archive
<i>Note.</i> All Warne coefficients are taken from the Angrist et al. (2021) HLO row in Table 1 of Warne (2023). The corresponding pairwise sample sizes appear in the upper triangle. The Henss rows use the archived variables identified in the table, not the direct-QNW, primary-HLO, or GHW variables on the Programme for International Student Assessment (PISA) scale used below. Country codes in Table 1 are KHM (Cambodia), CUB (Cuba), and PAK (Pakistan). SE denotes standard error. These entries concern source-specific numerical reproduction, not indicator equivalence.					

3 Data, constructions, populations, and samples

3.1 QNW and direct-QNW rule

QNW denotes the NIQ dataset's author-defined quality-and-recorded-sample-size-weighted national-IQ construction. The primary source is the NIQ-QNW field on the worksheet labelled FAV of *The NIQ-dataset* version 1.3.5, the October 2023 release (Becker 2023). The series is used here as a documented country-score compilation under audit, not as a consensus measure of national intelligence or an intrinsic characteristic of a population.

The primary rule is direct QNW v1.3.5. An economy is eligible when the QNW field is nonmissing, the NIQ metadata record psychometric data, and the estimate is not geographically imputed. This direct rule excludes the School Assessment Score (SAS) and geographic completion. It does not certify national representativeness, resolve source-study comparability, or make author-supplied quality ratings and recorded sample sizes independent assessments of validity. QNW v1.3.3, QNW + SAS, and QNW + SAS+GEO are retained only as named sensitivity constructions.

SAS is the NIQ School Assessment Score. Incorporating it makes QNW + SAS partly dependent on school-assessment information, thereby mechanically moving it toward learning indicators. QNW + SAS+GEO further permits geographic completion for economies with no direct observation available. Results for these variants cannot be interpreted as independent validation of HLO, nor can they be described as direct-QNW results.

3.2 HLO construction

HLO is the authors' documented primary aggregate reconstructed from the World Bank Harmonized Learning Outcomes source associated with Angrist et al. (2021). It is constructed from harmonized performance records for students assessed in international and regional testing programmes. It is not a census of residents, not a direct national psychometric estimate, and not a gold standard for QNW. The selected primary construction is distinct from the released country aggregate produced by the source repository's record-level mean. The latter is retained only as a replication benchmark.

The primary HLO estimand is mean learning performance over 2000–2017. The eligible standard components are mathematics, reading, and science at primary and secondary schooling levels. Within an economy, records are first averaged across years within each subject-by-level component. Subjects are then weighted equally within each level, and the primary and secondary level means are weighted equally. This subject-then-level hierarchy prevents economies with many records in one component from receiving greater weight merely because of record count.

Primary eligibility requires both schooling levels and at least two distinct subjects. The rule retained 82 economies before intersection with QNW or GDP. It was selected in HLO-only construction work on construct coverage, component comparability, temporal composition, and country coverage, rather than by optimizing an association with QNW or GDP. It prevents a single school level or a single domain from representing a general-learning composite while avoiding the much narrower all-six-component requirement.

Three HLO sensitivities are reported. *Latest HLO* selects the most recent observation for each component while retaining the primary eligibility rule. *Balanced all-six HLO* retains only economies with all three subjects at both levels and uses the 2000–2017 equal-weight hierarchy. *All available HLO* uses the published-period HLO score for every economy with a nonmissing value, without the primary component-eligibility gate. The latter changes coverage as well as the interpretation of the aggregate.

3.3 GHW comparator

GHW is the country mean mathematics-and-science achievement score reported in column 3 of Table A4 in Gust et al. (2024), on the PISA scale. Every economy in the frozen 66-economy primary ranking sample was checked against Table A4 and retained only when column 1 classified it as a Layer 1 or Layer 2 estimate. These are test-based, non-imputed estimates. Layer 2 values are transformed from the Trends in International Mathematics and Science Study (TIMSS) to the PISA scale. GHW differs from HLO in linking approach, assessment composition, grade or population, timing, and aggregation. Its inclusion broadens the descriptive comparison but does not create a clean experimental contrast between measurement families.

3.4 GDP, population, and exact samples

The GDP outcome is World Development Indicators (WDI) series NY.GDP.PCAP.PP.KD, GDP per capita at purchasing power parity (PPP), in constant 2021 international dollars (World Bank 2026a). A frozen World Bank extract was retrieved on 20 July 2026. The study uses the observed 2017 value and analyzes its natural logarithm, $\lambda_i = \log(\text{GDPpci})$. This is a contemporaneous country-level association, not a causal development model.

The full frozen World Bank frame contains 217 non-aggregate economies (World Bank 2026c). This frame must be distinguished from the HLO-conditioned frame: the HLO source has 164 records, of which 163 remain after excluding the historical Serbia and Montenegro (SCG) record. Of those 163, 162 match the 217-economy World Bank frame. Taiwan (TWN) is the one HLO reporting-frame record outside that universe. Within the strict World Bank frame, the flow is 217 economies, 162 with a matched HLO record (74.7%), 81 meeting the primary HLO rule (37.3% of 217 and 50.0% of the prior stage), 65 also having direct QNW (30.0% of 217 and 80.2% of the prior stage), and the same 65 having positive 2017 GDP. In the analytical HLO-conditioned frame, Taiwan increases the corresponding HLO, primary-HLO, and ranking counts to 163, 82, and 66. It lacks the frozen World Bank GDP value, so the exact GDP sample remains 65. These are statistical-frame statements rather than sovereignty claims.

An exact-sample comparison holds the economies fixed across all indicators in that comparison. The 66-economy ranking sample and the 65-economy GDP sample follow from the eligibility and matching rules described above. The GDP sample is selective. It contains 30.0% of the 217-economy frame. Under the frozen World Bank income classification, 64.6% of included economies are high income compared with 28.9% of excluded economies.

Standardized mean differences are 1.17 for log GDP and .78 for log population. China, India, Pakistan, Nigeria, and Bangladesh are among the populous economies absent from the GDP sample. The 65 included economies contain 33.9% of the full frame's 2017 population. Results therefore describe a selected, disproportionately affluent subset and are not population-representative global estimates.

4 Audit outputs and analytical methods

4.1 Outputs, estimands, and descriptive status

The audit distinguishes four outputs. First, Pearson, Spearman, and Kendall *tau*-b statistics summarize score association and ordinal agreement. Second, point-ranking sensitivity concerns ranks of the reported published scores in the exact common sample. Third, classification sensitivity concerns membership in the highest and lowest *k* economies. Fourth, downstream-association sensitivity concerns the bivariate association between each score and log GDP per capita on the same country set. None of these outputs identifies a common latent variable, a true country rank, or a causal relationship.

The rank estimand is deliberately narrow: descending average ranks of published country point scores. For economy *i* and indicators A and B in a sample of size *N*, absolute displacement is $|\text{rank}(A_i) - \text{rank}(B_i)|$. Its normalized form is

$$d_i = \frac{|\text{rank}(A_i) - \text{rank}(B_i)|}{N - 1}.$$

Normalization is required because an identical raw movement has a different meaning in samples of 42, 66, and 112 economies. Mean, median, maximum, signed, and absolute movements are descriptive attributes of the published point-score ordering. Comparable country-level uncertainty distributions are unavailable across the sources, so the article neither constructs ad hoc rank confidence intervals nor claims statistical reliability of individual rank changes.

The analysis-plan primary classification uses the 10th- and 90th-percentile score thresholds calculated separately for each indicator. This is the Hyndman–Fan/R type-7 convention: for ordered scores and probability *p*, ($h = 1 + (N-1)p$), with linear interpolation between the observations bracketing *h* (Hyndman and Fan 1996). Economies equal to a threshold are retained, so boundary ties can enlarge a tail. At *N* = 66, *h* = 7.5 for *p* = .10 and 59.5 for *p* = .90. In the absence of boundary ties, seven economies enter each tail. At *N* = 66, standard sample-quantile definitions need not produce the same threshold or realized tail size. Across the nine common Hyndman–Fan conventions in these untied data, realized tail counts range from six to eight. The quantile convention is therefore part of the classification definition. Point ranks use average ranks for tied scores. In the realized sample, neither QNW nor HLO contains score ties, so the planned result coincides numerically with *k* = 7. Fixed-*k* results at *k* = 5, 10, and 15 and the full *k* = 5–15 membership curve were added post hoc as cutoff-sensitivity analyses. Those fixed-*k* classifications contain exactly *k* economies, with scores sorted descending or ascending and any boundary tie resolved by ascending three-letter ISO 3166-1 country code (ISO3). That tie-break rule had no numerical effect in these data. For every pair and tail, intersection, union, and Jaccard similarity, $|A \cap B|/|A \cup B|$, are reported. No cutoff is treated as a general standard of acceptable substitution.

4.2 Score and ranking analyses

Pearson correlation describes linear score association. Spearman correlation and Kendall's *tau*-b summarize ordinal association. These statistics are reported alongside rank displacement because high aggregate association can coexist with consequential movement for particular economies. Each pairwise comparison is calculated on the same 66-economy three-way sample in the primary analysis. The analysis reports all pairwise results rather than treating QNW–HLO as a standalone contrast.

The GDP association is the Pearson correlation between a skill indicator and log 2017 GDP per capita. In a bivariate regression of standardized GDP on a standardized indicator, the slope equals that correlation. This identity is used only to make the reported slope contrast interpretable. It does not turn a bivariate ecological regression into a causal model.

4.3 Planned GDP inference and dependent-correlation comparison

For the common 65-economy GDP sample, the primary models regress sample-standardized log GDP on a sample-standardized indicator in separate bivariate equations. The slope therefore equals the corresponding Pearson correlation. Heteroskedasticity-consistent HC3 standard errors and confidence intervals are the primary regression inference (Long and Ervin 2000). A null-imposed Rademacher wild-bootstrap sensitivity based on 9,999 resamples was prespecified (Davidson and Flachaire 2008). Both models retain equal economy weights and the same 65 observations.

The two observed correlations and their difference are first descriptive, finite-sample quantities for this selected and effectively enumerated set of 65 economies (Abadie et al. 2020). The reported intervals and tests acquire an inferential target only under a hypothetical repeated-sampling or superpopulation interpretation, in which comparable economies are treated as independent observational units generated by a stable model. Under that interpretation, the target contrast is the superpopulation quantity $\Delta\rho = \rho(\text{QNW}, Y) - \rho(\text{HLO}, Y)$. HC3 allows regression-error variance to differ across economies, but it does not address spatial, regional, or other cross-economy dependence. The wild bootstrap uses the same independence structure. Williams' test and Zou's interval likewise require independent observational units and the usual regularity assumptions for a trivariate correlation model, including the multivariate-normal/reference-distribution approximation used by the Williams procedure and the Fisher-transformation approximation used by Zou's procedure. None of HC3, the bootstrap, Williams, or Zou repairs the nonrandom selection that produced the 65-economy sample. The model-based results are therefore reported as conventional conditional summaries, not design-based uncertainty for the enumerated finite set or globally representative inference.

The planned target contrast is $\Delta = r_{\text{QNW}, Y} - r_{\text{HLO}, Y}$, where both correlations overlap in the GDP variable and are estimated on the same economies. The standard Williams test (Williams 1959; see Steiger 1980, Eq. 7) evaluates equality of two overlapping correlations while retaining the dependence induced by the shared outcome and the QNW–HLO correlation. Hittner et al. (2003) examined the finite-sample performance of dependent-correlation tests. Zou's method supplies the corresponding confidence interval for the difference between overlapping correlations (Zou 2007). Neither a non-significant Williams test nor an interval containing zero establishes equivalence because no equivalence bound was prespecified.

The planned diagnostic set includes leverage, internally studentized residuals, Cook's distance, leave-one-economy estimates, and leave-one-region estimates. These diagnostics describe influence and regional sensitivity. They do not create an automatic deletion rule.

4.4 Construction, coverage, and weighting sensitivities

The sensitivity inventory is reported in three groups. Fixed-sample construction sensitivities preserve the primary exact sample while changing an indicator edition or HLO timing rule. Coverage-changing sensitivities alter the indicator definition, component eligibility, or population and therefore change both the sample and the estimand. Weighting sensitivities retain the 65 economies but change the target from the equal-economy-weighted correlation in the included sample to a population-weighted association.

The population-weighted correlation uses positive 2017 WDI Population, total series SP.POP.TOTL as weights (World Bank 2026b). For variables x and y , it uses weighted means, weighted covariance, and weighted standard deviations,

$r_w = \text{Cov}_w(x, y) / (\text{SD}_w(x) \text{SD}_w(y))$. It applies only to the selected 65-economy GDP sample. It neither restores absent economies nor corrects selection. For transparency, concentration of the population weights is summarized using the Kish weight-only index (Kish 1965), $n_w = (\sum w)^2 / \sum w^2 = 18.9$. This is a weight-concentration diagnostic, not an inferential sample size for the weighted correlation. It therefore answers a different, population-weighted question rather than supplying an alternative standard error for the equal-economy result.

The study was not preregistered. The analysis plan for RQ1 and RQ2 was finalized after data access and preliminary rank analysis. The fixed- k membership curves were added post hoc. Construction, coverage, weighting, and three-indicator comparisons are reported as sensitivity or extension analyses.

5 Results

5.1 Prior-claim audit

The source-specific results in Table 1 distinguish approximate or exact numerical reproduction of previously published results from the new analyses reported here. Calculations using the publisher supplement's National scores worksheet accompanying Warne (2023) reproduce the HLO–SAS (.966), HLO–QNW + SAS (.871), and HLO–geographic-only (.140) coefficients to three decimal places. They yield .809 rather than the published .811 for HLO–QNW and .830 rather than .829 for HLO–QNW + SAS+GEO. Sample sizes are larger by two or three economies in most reconstructed samples. Excluding Hong Kong, Macao, and Taiwan does not recover all published sample sizes and therefore does not fully explain the discrepancies.

By contrast, the coefficients reported in Table 2 of Henss (2024) are reproduced exactly at their displayed precision. After excluding Cambodia from NIQB and Cuba and Pakistan from HLO and GHW, the reconstructed coefficients are .841196 ($N = 158$) for NIQB–HLO IQ, .892595 ($N = 153$) for NIQB–GHW IQ, and .928059 ($N = 148$) for HLO IQ–GHW IQ. These values round to the reported .84, .89, and .93, respectively. This reproduces the displayed coefficients, but the NIQB definition includes SAS and geographic completion and therefore does not reproduce direct-QNW results.

The Parra-Kirkegaard final-score archive contains 197 reported records, while its companion results table contains 198 unique three-letter country-code rows. The latter's mean standard error is 2.556, which rounds to 2.56 and therefore does not reproduce the paper's stated 2.58 at two decimals. On the frozen 66-economy ranking sample, its score correlates .872 with QNW, .964 with HLO, and .964 with GHW. These high correlations are not used as independent validation because the composite incorporates cognate IQ and learning inputs. The article's GDP $r = .82$ is not reproduced because its four-source 2018–2022 GDP average was absent from the public repository files available on the audit date.

5.2 Score association and point-ranking sensitivity

Table 2 reports all three pairwise comparisons on the same 66 economies. QNW is similarly associated with HLO and GHW (Pearson $r = .795$ and $.793$), while the two learning indicators are more closely associated ($r = .954$). The corresponding rank results show the same broad pattern. QNW–HLO has mean absolute movement 9.52 places, median 7, maximum 34, and mean normalized movement .146. QNW–GHW has mean normalized movement .135. HLO–GHW has mean normalized movement .065.

Table 2
Pairwise exact-sample score and point-rank comparison (N = 66)

Comparison	Pearson <i>r</i>	Spearman <i>rho</i>	Kendall <i>tau-b</i>	Mean movement	Median	Maximum	Mean normalized movement
HLO–QNW	.795	.791	.597	9.52	7.0	34	.146
GHW–QNW	.793	.811	.625	8.79	6.5	33	.135
GHW–HLO	.954	.957	.827	4.20	3.0	15	.065

The result does not imply that a normalized displacement of .146 is universally material. Mean country-level absolute rank displacement divided by ($N-1$) was .146 in this 66-economy published-point-score sample. HLO and GHW agree more closely, but this difference is descriptive and cannot identify a single cause because their assessment environments overlap.

5.3 Extreme-group classification sensitivity

The analysis-plan empirical-decile rule retained threshold ties and used the type-7 linear quantile convention specified in Section 4.1. In the realized QNW–HLO data, there were no score or boundary ties. Each empirical tail contained seven economies, so the result equals the $k = 7$ column in Table 3. QNW and HLO shared four upper-tail economies (union = 10, Jaccard = .400) and two lower-tail economies (union = 12, Jaccard = .167). The $k = 5, 10$, and 15 columns are post hoc cutoff-sensitivity analyses, not predeclared cutoffs. Across these cutoffs, QNW–HLO overlap ranges from four of five to nine of fifteen economies in the upper tail and from two of five to ten of fifteen in the lower tail. GHW–QNW shows a similar pattern, whereas GHW–HLO overlaps more. These are classifications of published point scores, not probabilistic membership or demonstrated decision error.

Table 3
Highest and lowest group overlap in the 66-economy exact sample

Pair	Tail	$k = 5$: overlap/Jaccard	$k = 7$: overlap/Jaccard	$k = 10$: overlap/Jaccard	$k = 15$: overlap/Jaccard
HLO–QNW	Highest	4/.67	4/.40	5/.33	9/.43
HLO–QNW	Lowest	2/.25	2/.17	4/.25	10/.50
GHW–QNW	Highest	4/.67	4/.40	5/.33	9/.43
GHW–QNW	Lowest	1/.11	2/.17	5/.33	9/.43
GHW–HLO	Highest	5/1.00	6/.75	8/.67	11/.58
GHW–HLO	Lowest	3/.43	5/.56	9/.82	13/.76

Figure 1 QNW–HLO top- and bottom- k Jaccard similarity for $k = 5–15$ in the fixed 66-economy sample. Each classification contains exactly k economies. The ISO3 boundary rule had no effect because the relevant scores were untied. This post hoc analysis describes sensitivity to the cutoff, not rank uncertainty

5.4 Planned GDP associations, inference, and influence diagnostics

Among the 65 economies with both focal indicators and 2017 GDP per capita, the QNW–GDP correlation is .616 and the HLO–GDP correlation is .690. In the corresponding standardized simple regressions, HC3 standard errors are .091 and .101 and the null-imposed wild-bootstrap p values are .0002 and .0001, respectively (Table 4). These coefficients describe the selected equal-economy sample. They do not identify causal effects or population-representative global associations.

Table 4
Planned GDP regressions and dependent-correlation comparison (N = 65)

Indicator	Standardized coefficient (= Pearson r)	HC3 SE	HC3 95% CI	HC3 p	R^2	Wild-bootstrap p
QNW	.615905	.091413	[.433230, .798580]	5.72e-9	.379339	.0002
HLO	.689519	.100889	[.487909, .891129]	3.88e-9	.475437	.0001
<i>Note.</i> The indicators correlate .789646 with each other. The planned overlapping-correlation contrast is $\Delta r = r_{QNW,GDP} - r_{HLO,GDP} = -.073614$. The Williams test gives $t(62)=-1.243932$, $p=.218207$. The Zou 95% CI is [-.207076, .045079]. Wild-bootstrap values use the prespecified procedure described in Section 4.3. The observed correlations are descriptive for the selected finite set. The intervals and tests are conventional model-based results under the assumptions stated in Section 4.3.						

The point estimate favors a larger HLO–GDP correlation, but the planned Williams test does not reject equality and the Zou interval includes zero and differences in either direction. This is uncertainty, not evidence of equivalence. The GHW–GDP correlation of .748 is a descriptive extension and is not part of the planned focal contrast.

The planned diagnostics identify influence without supplying an automatic deletion rule. Ghana has the largest leverage (.190) in the QNW regression. Qatar has the largest absolute internally studentized residual (3.776) and Cook's distance (.294), both in the HLO regression. Across all 65 leave-one-economy comparisons, the QNW-minus-HLO contrast remains negative, ranging from $-.103$ when Qatar is omitted to $-.035$ when the Philippines is omitted. It also remains negative after every leave-one-region deletion. Omitting East Asia and the Pacific nearly eliminates it, however: $N = 55$, $r_{QNW} = .638782$, $r_{HLO} = .639254$, and $\Delta r = -.000471$. Thus direction survives the recorded deletions, but magnitude is region-sensitive.

5.5 Construction, coverage, and weighting sensitivity pattern

Table 5 contains fixed-sample construction sensitivities. The QNW v1.3.3 and latest-HLO rows retain the baseline ranking and GDP sample sizes, so their movements can be compared after normalization. They are still alternative constructions rather than estimates of sampling error. Latest HLO increases the QNW–HLO mean normalized movement from .146 to .164 and changes the HLO GDP association from .690 to .710.

Table 5
Fixed-sample construction sensitivities

Scenario	Exact construction change	Ranking N	GDP N	QNW–HLO mean normalized movement	QNW–GDP <i>r</i>	HLO–GDP <i>r</i>	Difference
Primary reference	Direct QNW v1.3.5 with primary 2000–2017 HLO	66	65	.146	.616	.690	– .074
Earlier QNW	Direct QNW v1.3.3 with the same primary HLO rule	66	65	.149	.607	.690	– .082
Latest HLO	Direct QNW v1.3.5 with the latest component observation and primary eligibility	66	65	.164	.616	.710	– .094

Table 6 contains scenarios that change coverage, the indicator definition, or both. QNW + SAS and QNW + SAS+GEO contain school-assessment information. Their small rank movements and + .009 GDP contrasts are therefore not sign reversals of the same direct-QNW estimand or independent corroboration of agreement with HLO. The GEO extension adds 28 national-IQ records to the broader source pool, but none satisfies the primary-HLO eligibility rule. Consequently, QNW + SAS and QNW + SAS+GEO use the same 81 economies for ranking, the same 80 for GDP, and identical scores on those intersections. Their equal results are mechanical, not independent replications. Balanced all-six HLO changes component comparability and coverage, all-available HLO removes the primary component gate, and the UN row changes the statistical population. Each row has its own exact economy set.

Table 6
Coverage-changing indicator and population sensitivities

Scenario	What changes	Ranking N	GDP N	QNW–HLO mean normalized movement	QNW–GDP <i>r</i>	HLO–GDP <i>r</i>	Difference
UN member/observer population	Statistical population	64	64	.154	.607	.683	– .076
QNW + SAS	Indicator and coverage, including school-assessment information	81	80	.062	.718	.709	.009
QNW + SAS+GEO	Indicator and coverage, also permitting geographic completion	81	80	.062	.718	.709	.009
Balanced all-six HLO	HLO construction and coverage	42	41	.153	.567	.601	– .034
All available HLO	Removes primary HLO component-eligibility gate	112	107	.134	.757	.762	– .005

Differences in Tables 5–7 are computed from unrounded correlations and then displayed to three decimals. Within fixed-sample comparisons using direct QNW, the HLO–GDP point correlation is consistently larger: the QNW-minus-HLO contrast is $-.074$ in the primary analysis, $-.082$ with QNW v1.3.3, and $-.094$ with latest HLO. Population weighting gives $-.133$, and every leave-one-economy estimate remains negative. The UN-frame result is $-.076$ and therefore is not attenuated relative to the primary estimate. Coverage-changing indicator-construction analyses do attenuate the contrast: balanced and all-available estimates are $-.034$ and $-.005$, while the school-assessment-augmented constructions yield $+.009$. Overall, point-estimate direction is stable under the recorded direct-QNW fixed-sample perturbations, but magnitude depends materially on construction, coverage, weighting, and regional composition. The planned Williams test and Zou interval do not rule out a zero contrast.

Table 7 changes only weighting within the selected sample, not country coverage. The equal-economy result is the equal-economy-weighted correlation in the included sample. With 2017 population weights, the QNW and HLO correlations are $.734$ and $.867$. This weighting does not add China, India, Pakistan, Nigeria, Bangladesh, or any other omitted economy. Its Kish weight-only index is 18.9. The larger difference therefore belongs to a concentrated, population-weighted estimand for the same 65 selected economies and is not a replacement for Table 4's equal-economy comparison.

Table 7
Weighting and estimand sensitivity in the fixed 65-economy GDP sample

Estimand	Weight	N	QNW–GDP r	HLO–GDP r	Difference
Equal-economy association	Each economy weight 1	65	$.616$	$.690$	$-.074$
Population-weighted association	WDI 2017 total population	65	$.734$	$.867$	$-.133$

6 Discussion

The primary result is narrow. In the stated 66-economy exact sample, direct QNW has high aggregate association with both learning indicators, but high association alone does not determine the point ranks or highest/lowest group memberships reported here. HLO and GHW agree more closely with one another than either agrees with QNW, while still sharing a partly overlapping assessment ecosystem. The result does not show that learning indicators are universally preferable, that national-IQ series are invalid, or that a single construct difference explains the observed movements.

The GDP analysis does not establish superiority or equivalence. HLO's point correlation is larger in the primary sample, in both direct-QNW fixed-sample variants, under population weighting, and after every recorded economy or region deletion. Yet the planned Williams test is non-significant, and the Zou interval spans zero. Omitting East Asia and the Pacific reduces the point contrast almost to zero. Coverage-changing indicator-construction analyses attenuate it or yield $+.009$, whereas the UN-frame result does not. The exact pattern is therefore neither “HLO always wins” nor “the direction is generally unstable.” The QNW-minus-HLO point contrast remains negative across the direct-QNW fixed-sample analyses, while model-based uncertainty and estimand-changing sensitivities limit the inference. A proposed substitution must identify its population, output, and tolerance before any aggregate correlation is treated as sufficient.

The prior-claim audit provides a related lesson. Warne, Henss, and Parra–Kirkegaard involve different indicators, transformations, source versions, and coverage. Henss's correlations can be reproduced only when NIQB is correctly identified as QNW + SAS+GEO and the stated exceptional values are excluded. Calling a direct-QNW computation a

reproduction would have been a source-definition error. Reproducibility work is most informative when it preserves such boundaries rather than forcing unlike measures into a common numerical table.

Taken together, these findings have implications beyond the specific country-level indicators examined here. Similar problems may arise whenever one measure is replaced by another in research or decision-making, including in education, health, governance, environmental assessment, and financial risk. A clear substitution study states what the new measure is expected to preserve, such as scores, rankings, classifications, associations, or decisions. A high correlation between two measures does not guarantee stability in these other outputs. Using the same observations helps isolate the effect of replacing the measure. Data sources, sample rules, and exclusions also need to be clear. Where a defensible tolerance exists, it can be stated in advance. Otherwise, results are best treated as descriptive. Changes in construction are best separated from changes in coverage, eligibility, imputation, or weighting because these choices alter both the sample and the question being answered. Reproducibility also depends on documenting the data, code, and exact analytical samples (Stodden et al. 2018; Hardwicke et al. 2018). These are practical reporting considerations, not a formally validated checklist.

7 Limitations

The analysis has several limits. First, country scores and ranks are published point estimates, and comparable uncertainty models are unavailable across sources. Second, only 65 of 217 economies in the frozen World Bank frame enter the GDP analysis. Major populous economies are absent, and the population weights are highly concentrated among the included economies, with a Kish weight-only index of 18.9. The finite-set correlations are descriptive. The reported intervals and tests depend on an unverified superpopulation or repeated-sampling interpretation, independence across economies, and method-specific model assumptions. HC3 does not address cross-economy dependence, and none of the procedures corrects selection. Third, period, assessed grade or population, source tests, linking, aggregation, and coverage vary together. Differences therefore cannot be attributed to a single feature or to an IQ-versus-learning distinction. Fourth, no decision loss or equivalence margin was prespecified, so the paper cannot determine whether the observed output differences are scientifically tolerable. Fifth, GHW is separately constructed but is not independent of the broader assessment environment. Finally, all analyses are ecological and descriptive. They neither imply individual-level relationships nor identify causal effects on development.

8 Conclusion

High cross-country correlation alone does not establish that country-level indicators are interchangeable for particular analytical purposes. In the fixed samples examined here, replacing direct QNW with HLO altered point ranks and extreme-group membership despite strong score association. HLO also had a larger point correlation with GDP, although the dependent-correlation inference and sensitivity analyses did not establish a stable or general advantage. These findings apply only to the specified indicators, constructions, and selected economy samples. They neither validate nor invalidate entire measurement families. Future substitution claims should identify the relevant population, output, and tolerance and distinguish changes in indicator construction from changes in coverage or weighting. The broader lesson is that indicator substitution cannot be judged by correlation alone.

Declarations

Author Contribution

G.K. conceived and designed the study, curated the data, developed the methodology and analysis code, conducted the formal analysis, prepared the figures, wrote and revised the manuscript and approved the final version.

Data Availability

This study uses publicly available secondary data obtained from 1. the NIQ Dataset v1.3.5 (<https://viewoniq.org/wp-content/uploads/2023/10/NIQ-DATASET-V1.3.5.zip>), 2. the World Bank Harmonized Learning Outcomes database (https://datacatalogfiles.worldbank.org/ddh-published/0038001/1/DR0046131/hlo_database.xlsx) (https://datacatalogfiles.worldbank.org/ddh-published/0038001/1/DR0046131/hlo_database.xlsx), 3. World Development Indicators GDP per capita series NY.GDP.PCAP.PP.KD (<https://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.PP.KD?downloadformat=csv>) (<https://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.PP.KD?downloadformat=csv>), 4. World Development Indicators population series SP.POP.TOTL (<https://api.worldbank.org/v2/en/indicator/SP.POP.TOTL?downloadformat=csv>) (<https://api.worldbank.org/v2/en/indicator/SP.POP.TOTL?downloadformat=csv>), and 5. Gust–Hanushek–Woessmann comparator (<https://hanushek.stanford.edu/sites/default/files/Gust%20Hanushek%20Woessmann%202024%20JDE%20166.pdf>) (<https://hanushek.stanford.edu/sites/default/files/Gust%20Hanushek%20Woessmann%202024%20JDE%20166.pdf>)).

References

1. Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M.: Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*. **88**(1), 265–296 (2020). <https://doi.org/10.3982/ECTA12675>
2. Angrist, N., Djankov, S., Goldberg, P.K., Patrinos, H.A.: Measuring human capital using global learning data. *Nature*. **592**(7854), 403–408 (2021). <https://doi.org/10.1038/s41586-021-03323-7>
3. Becker, D.: The NIQ-dataset, version 1.3.5 [Data set] (2023). https://viewoniq.org/?page_id=9. Accessed 20 July 2026
4. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. **327**(8476), 307–310 (1986). [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)
5. Bosker, R.J.: The different levels of the g-factor. *Eur. J. Pers.* **21**(5), 712–714 (2007). <https://doi.org/10.1002/per.648>
6. Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**(1), 37–46 (1960). <https://doi.org/10.1177/001316446002000104>
7. Davidson, R., Flachaire, E.: The wild bootstrap, tamed at last. *J. Econ.* **146**(1), 162–169 (2008). <https://doi.org/10.1016/j.jeconom.2008.08.003>
8. Goldstein, H., Spiegelhalter, D.J.: League tables and their limitations: statistical issues in comparisons of institutional performance. *J. Royal Stat. Society: Ser. A*. **159**(3), 385–409 (1996). <https://doi.org/10.2307/2983325>
9. Gust, S., Hanushek, E.A., Woessmann, L.: Global universal basic skills: current deficits and implications for world development. *J. Dev. Econ.* **166**, 103205 (2024). <https://doi.org/10.1016/j.jdeveco.2023.103205>
10. Hall, P., Miller, H.: Using the bootstrap to quantify the authority of an empirical ranking. *Annals Stat.* **37**(6B), 3929–3959 (2009). <https://doi.org/10.1214/09-AOS699>
11. Hardwicke, T.E., Mathur, M.B., MacDonald, K., Nilsonne, G., Banks, G.C., Kidwell, M.C., Hofelich Mohr, A., Clayton, E., Yoon, E.J., Tessler, M.H., Lenne, R.L., Altman, S., Long, B., Frank, M.C.: Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Soc. Open. Sci.* **5**(8), 180448 (2018). <https://doi.org/10.1098/rsos.180448>

12. Henss, R.: The intelligence of nations. National IQs and correlates, preprint v1. Qeios (2024).
<https://doi.org/10.32388/6IBSKW>
13. Hittner, J.B., May, K., Silver, N.C.: A Monte Carlo evaluation of tests for comparing dependent correlations. *J. Gen. Psychol.* **130**(2), 149–168 (2003). <https://doi.org/10.1080/00221300309601282>
14. Hunt, E., Wittmann, W.W.: National intelligence and national prosperity. *Intelligence.* **36**(1), 1–9 (2008).
<https://doi.org/10.1016/j.intell.2006.11.002>
15. Hyndman, R.J., Fan, Y.: Sample quantiles in statistical packages. *Am. Stat.* **50**(4), 361–365 (1996).
<https://doi.org/10.1080/00031305.1996.10473566>
16. Kendall, M.G.: The treatment of ties in ranking problems. *Biometrika.* **33**(3), 239–251 (1945).
<https://doi.org/10.1093/biomet/33.3.239>
17. Kish, L.: *Survey Sampling*. Wiley, New York (1965)
18. Lakens, D.: Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychol. Personality Sci.* **8**(4), 355–362 (2017). <https://doi.org/10.1177/1948550617697177>
19. Lin, L.I.-K.: A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* **45**(1), 255–268 (1989).
<https://doi.org/10.2307/2532051>
20. Long, J.S., Ervin, L.H.: Using heteroscedasticity consistent standard errors in the linear regression model. *Am. Stat.* **54**(3), 217–224 (2000). <https://doi.org/10.1080/00031305.2000.10474549>
21. Lynn, R., Meisenberg, G., Mikk, J., Williams, A.: National IQs predict differences in scholastic achievement in 67 countries. *J. Biosoc. Sci.* **39**(6), 861–874 (2007). <https://doi.org/10.1017/S0021932007001964>
22. Parra, L., Kirkegaard, E.O.W.: National IQs: measurement and defense. *OpenPsych* (2025).
<https://doi.org/10.26775/OP.2025.11.17>
23. Rindermann, H.: The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *Eur. J. Pers.* **21**(5), 667–706 (2007). <https://doi.org/10.1002/per.634>
24. Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **15**(6), 657–680 (1987).
<https://doi.org/10.1007/BF01068419>
25. Simonsohn, U., Simmons, J.P., Nelson, L.D.: Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208–1214 (2020).
<https://doi.org/10.1038/s41562-020-0912-z>
26. Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W.: Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**(5), 702–712 (2016). <https://doi.org/10.1177/1745691616658637>
27. Steiger, J.H.: Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87**(2), 245–251 (1980).
<https://doi.org/10.1037/0033-2909.87.2.245>
28. Stodden, V., Seiler, J., Ma, Z.: An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* **115**(11), 2584–2589 (2018).
<https://doi.org/10.1073/pnas.1708290115>
29. Warne, R.T.: National mean IQ estimates: validity, data quality, and recommendations. *Evolutionary Psychol. Sci.* **9**, 197–223 (2023). <https://doi.org/10.1007/s40806-022-00351-y>
30. Wicherts, J.M., Dolan, C.V., van der Maas, H.L.J.: A systematic literature review of the average IQ of sub-Saharan Africans. *Intelligence.* **38**(1), 1–20 (2010). <https://doi.org/10.1016/j.intell.2009.05.002>
31. Williams, E.J.: The comparison of regression variables. *J. Roy. Stat. Soc. B.* **21**(2), 396–399 (1959).
<https://doi.org/10.1111/j.2517-6161.1959.tb00346.x>

32. World Bank: GDP per capita, PPP (constant 2021 international \$), NY.GDP.PCAP.PP.KD. World Development Indicators [Data set and metadata]: (2026a). <https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD>. <https://databank.worldbank.org/metadataglossary/world-development-indicators/series/NY.GDP.PCAP.PP.KD>. Accessed 20 July 2026

33. World Bank: Population, total, SP.POP.TOTL. World Development Indicators [Data set]: (2026b). <https://data.worldbank.org/indicator/SP.POP.TOTL>. Accessed 20 July 2026

34. World Bank: World Bank Country and Lending Groups, FY2027 [Data set]: (2026c). https://api.worldbank.org/v2/country?format=json&per_page=400. Accessed 21 July 2026

35. Zou, G.Y.: Toward using confidence intervals to compare correlations. Psychol. Methods. **12**(4), 399–413 (2007). <https://doi.org/10.1037/1082-989X.12.4.399>

Figures

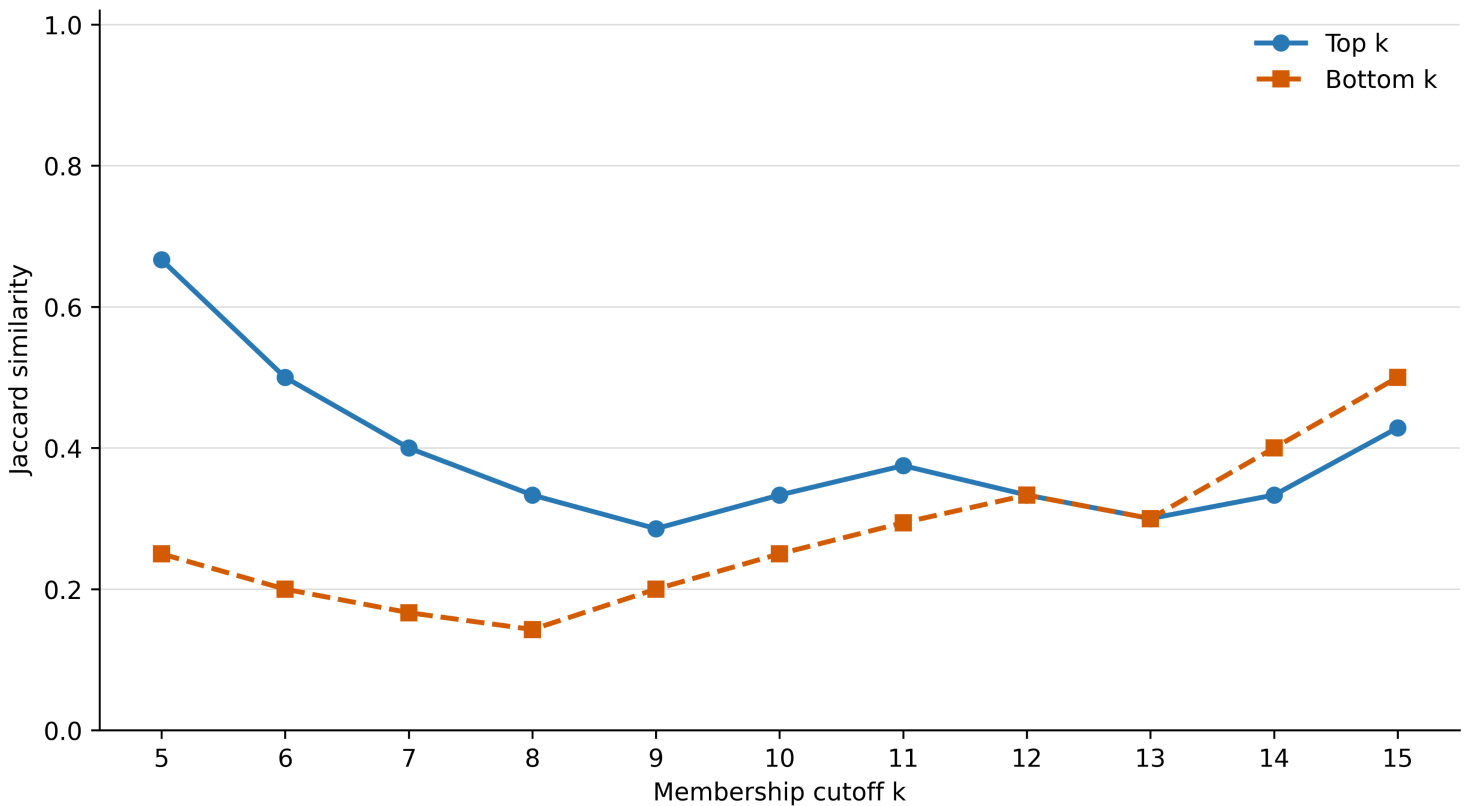


Figure 1

QNW–HLO top- and bottom-*k* Jaccard similarity for *k* = 5–15 in the fixed 66-economy sample. Each classification contains exactly *keconomies*. The ISO3 boundary rule had no effect because the relevant scores were untied. This post hoc analysis describes sensitivity to the cutoff, not rank uncertainty