

Grade 4-to-5 Joint Mathematics–Science Benchmark Transitions in Nine Systems: A Descriptive Common-Severity Analysis Using TIMSS 2023 Longitudinal Data

Jonas A. Mandalunes

jonas.mandalunes24414@oed.com.ph

AMA Computer University

Research Article

Keywords: international large-scale assessment, longitudinal achievement, mathematics, science, joint benchmark attainment, plausible values, jackknife repeated replication, standardization

Posted Date: August 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10682415/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Grade 4-to-5 Joint Mathematics–Science Benchmark Transitions in Nine Systems: A Descriptive Common-Severity Analysis Using TIMSS 2023 Longitudinal Data

1 Abstract

Background: Official mathematics and science means and benchmark shares describe domains separately and cannot recover joint longitudinal movement. We estimated the conditional transition from below both benchmarks (BB) to at or above both (AA) and examined its dependence on baseline-severity distributions.

Methods: We analyzed TIMSS 2023 Longitudinal Grade 4-to-5 data for 44,325 students in nine systems and, separately, Grade 8-to-9 data for 16,764 students in three systems. Four joint states at the primary 475-point threshold defined complete transition matrices and the BB-to-AA transition. Survey-weighted transition surfaces were integrated over a shared equal-system baseline-severity target, with five alternative specifications. Five matched four-dimensional plausible-value tuples and TIMSS jackknife repeated replication propagated measurement, sampling, shared-support, and shared-target uncertainty.

Results: Grade 4-to-5 BB-to-AA estimates ranged from 5.5% to 34.1%; common-severity estimates ranged from 8.6% to 25.6%, and standardized-minus-raw differences ranged from –9.8 to +6.2 percentage points. None of the 44 alternative-minus-primary 95% confidence intervals excluded zero, although the strict-common-cell intervals remained compatible with nontrivial differences. Changing the threshold changed both the risk set and event: estimates at 400 ranged from 14.0% to 47.0%. In a separate inverse-probability-weighted analysis targeting baseline participants, system-specific estimates differed from the PV-completed primary estimates by at

most 1.3 percentage points. In the separate three-system Grade 8-to-9 extension, BB-to-AA estimates ranged from 5.4% to 8.1%.

Conclusions: Joint transitions answer a conditional cross-domain question unavailable from marginal reporting. Common-severity standardization makes target dependence visible but does not identify causal effects, system effectiveness, or a pure composition component. Results are descriptive and specific to participating systems, thresholds, supports, and modeled targets.

Keywords: international large-scale assessment; longitudinal achievement; mathematics; science; joint benchmark attainment; plausible values; jackknife repeated replication; standardization

2 Background

International large-scale assessments usually report mean achievement, changes in means, and percentages reaching subject-specific benchmarks. These statistics answer important questions, but each treats mathematics and science separately. A student may improve in both domains while remaining below one benchmark, cross a benchmark in one domain while losing ground in the other, or move from below both benchmarks to simultaneous attainment. None of these pathways can be reconstructed from two sets of marginal percentages alone. A joint longitudinal representation is needed when the substantive question concerns coordinated movement across domains rather than isolated domain growth.

TIMSS 2023 Longitudinal followed students assessed in 2023 into an additional assessment in 2024. The official database, reports, and current overview appropriately emphasize population means, benchmark distributions, domain-specific growth, and the database's new opportunities for longitudinal secondary analysis (Davies et al., 2025; Fishbein, Taneva, et al., 2026; Fishbein, Bezirhan, et al., 2026). Secondary analyses can complement those summaries by asking how

students move among *joint* mathematics–science states. Such analyses must preserve the TIMSS measurement and sampling architecture: achievement is represented by plausible values rather than individual test scores, and uncertainty must include both imputation and complex-sample components (Fishbein et al., 2024; Mislevy et al., 1992; Rutkowski et al., 2010; Wu, 2005). Recent guidance likewise emphasizes using all plausible values for population inference and propagating the relevant design uncertainty (Huang, 2024; Jewsbury et al., 2024).

A second issue arises when conditional joint transitions are compared across education systems. Students classified below both benchmarks need not begin equally far below them. A system whose BB risk set lies close to the two cut scores offers more opportunity for threshold crossing than a system whose risk set begins farther away. A raw BB-to-AA probability therefore combines the observed baseline mathematics–science severity distribution with the conditional relationship between baseline severity and follow-up joint attainment. A model-based common-severity analysis can show how estimates change under a shared target, but it cannot cleanly separate “composition” from “performance” or identify why conditional relationships differ.

Transition descriptions, benchmark classification, direct standardization, plausible-value inference, and jackknife repeated replication each have established uses (Fishbein et al., 2024; Keiding & Clayton, 2014; Lane & Nelder, 1982; Mislevy et al., 1992; Rutkowski et al., 2010; Wu, 2005). Adjacent research has examined joint mathematics–science profiles and the broader primary-to-secondary transition literature (Berger et al., 2020; Kaur et al., 2022), but those studies address different constructs and estimands. We conducted a focused, nonsystematic search of TIMSS, large-scale-assessment, and methodological literature to position this analysis; that search does not establish priority. We integrate established components in a descriptive application comprising a four-state cross-domain longitudinal estimand, a BB-to-AA transition

to joint attainment, shared baseline-severity standardization, matched four-dimensional longitudinal/cross-domain plausible values, and replicate-wise rebuilding of shared supports and targets with full covariance. The contribution is the coordinated application and transparent inference workflow, not a claim that any component or estimator is individually new.

The analysis is descriptive and deliberately avoids causal interpretation. Education systems were not randomized, measured system characteristics are not exchangeable treatments, and a one-year transition cannot diagnose individual learning mechanisms. The estimands describe benchmark-state movement among participating systems and show how that description changes under specified common-severity targets. They are not system rankings, evaluations of instructional quality, estimates of pure distributional composition, or estimates for a global population.

3 Conceptual framework and research questions

Let mathematics and science plausible values at baseline be M_0 and S_0 , and at follow-up be M_1 and S_1 . For threshold c , each wave is represented by one of four mutually exclusive states:

- **BB:** $M < c$ and $S < c$, below both benchmarks;
- **MB:** $M < c$ and $S \geq c$, mathematics below only;
- **SB:** $M \geq c$ and $S < c$, science below only; and
- **AA:** $M \geq c$ and $S \geq c$, at or above both benchmarks.

The resulting 4×4 matrix distinguishes persistence, one-domain movement, joint attainment, and movement out of attainment. We use “benchmark attainment,” not “proficiency” or “passing,” because TIMSS international benchmarks are descriptive scale points, not diagnostic

standards for individual students (Davies et al., 2024). BB denotes a baseline state, not evidence that a student had previously attained either benchmark. Accordingly, the focal quantity is called the **BB-to-AA transition to joint attainment**, or **BB-to-AA transition**.

The primary transition is

$$\theta_s^{\text{raw}} = P_w(Z_1 = AA \mid Z_0 = BB, \text{ system} = s),$$

where P_w denotes a population-weighted probability and Z_t is the joint state at wave t . A student must begin below 475 in both domains and later be at or above 475 in both.

We address four research questions:

1. **Joint movement and attainment:** What are the Grade 4-to-5 four-state mathematics–science transition matrices, and what proportion of the baseline BB risk set makes the BB-to-AA transition to joint attainment?
2. **Common-severity target:** How do system-specific BB-to-AA estimates change when modeled transitions are integrated over the same distribution of baseline mathematics and science severity?
3. **Sensitivity of the primary finding:** How do the primary BB-to-AA estimates and common-severity conclusions change across prespecified thresholds, support and response-surface alternatives, participation and missingness analyses, and the exploratory same-wave classification-discordance diagnostic?
4. **Descriptive grade-band extension:** What joint movement and attainment patterns appear in the distinct Grade 8-to-9 longitudinal samples for Jordan, the Republic of Korea, and Sweden?

4 Methods

4.1 Data, systems, and analytic samples

We used the official TIMSS 2023 Longitudinal international database, user guide, technical report, and sampling documentation (Fishbein et al., 2025; Fishbein, Taneva, et al., 2026). The Grade 4-to-5 analytic file contained 44,325 students from Georgia, Italy, Jordan, the Republic of Korea, Kosovo, Montenegro, North Macedonia, Slovenia, and Sweden. The separate Grade 8-to-9 file contained 16,764 students from Jordan, the Republic of Korea, and Sweden. Grade bands were not pooled.

The analytic files retained student and school identifiers, final student weight (TOTWGT), jackknife zone and member indicators, five plausible values in each domain and wave, TIMSS-provided change variables, and wave-participation indicators. Fail-fast input gates required finite positive weights, finite zone and member values, JKREP in $\{0, 1\}$, both members represented within every system-zone, and process-participation codes in the observed documented domain $\{1, 2\}$; code 1 denotes participation. All achievement plausible-value cells required by the analyses were populated. Depending on the system, 81.3%–91.6% of the weighted Grade 4-to-5 sample participated in both waves; corresponding Grade 8-to-9 percentages were 81.6%–86.7% (Table 1). The public achievement files provide plausible values even where process indicators do not show participation at both waves. We therefore treat the primary analysis as the official PV-completed analytic representation and compare it explicitly with both-wave and response-weighted analyses.

132 **Table 1. Analytic samples and participation patterns**

Grade band	System	Students	Schools	JRR zones	Both-wave <i>n</i>	Baseline-only <i>n</i>	Follow-up-only <i>n</i>	Both-wave weighted %
Grade s 4–5	Georgia	4,491	156	79	3,787	624	80	84.3
Grade s 4–5	Italy	4,610	152	76	4,207	246	157	91.6
Grade s 4–5	Jordan	6,172	224	115	5,141	924	107	81.9
Grade s 4–5	Republic of Korea	4,507	156	78	3,558	797	152	81.3
Grade s 4–5	Kosovo	4,747	152	86	4,190	419	138	88.3
Grade s 4–5	Montenegro	4,559	141	125	4,150	267	142	90.8
Grade s 4–5	North Macedonia	4,866	150	87	4,137	505	224	85.2
Grade s 4–5	Slovenia	5,015	149	75	4,104	642	269	82.4
Grade s 4–5	Sweden	5,358	160	81	4,436	703	219	82.9
Grade s 8–9	Jordan	7,051	225	114	5,921	1,041	89	81.6
Grade s 8–9	Republic of Korea	4,410	176	88	3,845	412	153	86.7
Grade s 8–9	Sweden	5,303	154	78	4,358	724	221	82.5

133 *Note.* JRR = jackknife repeated replication. A wave is marked participated when either relevant
134 process flag equals 1; source codes were gated to {1, 2}. No students were classified as
135 participating in neither wave in the linked analytic files. Weighted percentages use the final
136 student weight (TOTWGT).

4.2 Threshold choice and transition quantities

The 475 Intermediate International Benchmark was designated as the primary threshold during analysis planning, before BB-to-AA estimates were calculated; the study was not externally preregistered. The decision was based on denominator feasibility. At 400, the dual-deficit risk set was very small in higher-performing systems (for example, a mean of approximately 26 students across plausible values in the Republic of Korea), whereas 475 produced a mean risk-set count of 162.8 in that system and 794.6–4,001.4 in the others. A higher threshold yields a larger baseline BB denominator and permits all nine systems to contribute to the same primary estimand. Thresholds 390, 400, 410, 465, 475, and 485 were retained as planned sensitivity analyses. This threshold choice does not redefine the official benchmark hierarchy.

For each plausible-value draw, we computed all 16 conditional transitions. Each transition-matrix row uses the weighted baseline-state total as its denominator. We also estimated baseline and follow-up state distributions, the probability that any baseline deficit state reached AA, BB persistence, and weighted changes among baseline-BB students in mathematics, science, and the weakest-domain score $\min(M, S)$. The continuous weakest-domain change complements threshold crossing by retaining distance information.

4.3 Common-severity standardization and six specifications

Within each plausible-value draw, we retained baseline-BB students and defined deficits

$$d_M = (475 - M_0)/100, \quad d_S = (475 - S_0)/100.$$

The primary specification fitted a survey-weighted, system-specific quadratic logistic joint-attainment surface:

$$\text{logit}\{P(Y = 1 \mid d_M, d_S, s)\} = \beta_{0s} + \beta_{1s}d_M + \beta_{2s}d_S + \beta_{3s}d_M^2 + \beta_{4s}d_S^2 + \beta_{5s}d_Md_S,$$

where $Y = 1$ denotes follow-up AA. Positive survey weights were divided by their system-specific mean before fitting. This constant rescaling leaves weighted score equations and coefficient estimates unchanged while avoiding numerical collapse in binomial initialization. For numerical stability, predicted probabilities used linear predictors clamped to $[-30, 30]$ before the inverse-logit transformation; diagnostics retained the unclamped range and counts and weighted shares of affected predictions.

Primary common support was the intersection of system-specific weighted 1st–99th percentile intervals in baseline mathematics and science. Within this rectangle, each system’s retained empirical distribution was weighted to mass 1/9, creating a synthetic equal-system target rather than a global-student target. The primary standardized estimate was

$$\theta_s^{\text{std}} = \sum_i q_i \text{logit}^{-1}(X_i \hat{\beta}_s),$$

where q_i is the common target weight. Models used each target system’s complete BB risk set; integration used the shared target. Retained weighted shares, effective sample sizes, target-system convex-hull coverage, and five-by-five joint-cell coverage qualify interpretation because a rectangular marginal intersection does not guarantee dense joint overlap and standardized comparisons require an explicit target and adequate support (Keiding & Clayton, 2014; Lane & Nelder, 1982; Westreich & Cole, 2010).

Six specifications were carried through one joint covariance analysis. The primary specification was retained unchanged. Five one-factor sensitivity alternatives were frozen internally before

their own estimates were examined, after the primary had been estimated. This chronology is an internal analysis-control record, not an external preregistration:

1. **Pooled represented-BB target (sensitivity alternative):** original student weights across retained baseline-BB records from all nine systems, rather than equal system mass; this is not a global target.
2. **Leave the Republic of Korea out (sensitivity alternative):** eight-system support and equal-system target, with no Korea estimate under this specification.
3. **Strict common cells (sensitivity alternative):** a replicate-specific 4×4 target-quartile grid retaining only cells with unweighted $n \geq 5$ and Kish effective sample size (ESS) ≥ 3 in every system.
4. **Additive linear surface (sensitivity alternative):** intercept plus linear mathematics and science deficits.
5. **Linear-interaction surface (sensitivity alternative):** intercept, both linear deficits, and their product, without squared terms.

For every system and applicable specification k , four quantities were retained: $\theta_s^{\text{raw full}}$, observed in the complete BB risk set; $\theta_{sk}^{\text{raw support}}$, observed in that specification's support; $\theta_{sk}^{\text{model own support}}$, the fitted surface averaged over the system's supported empirical distribution; and $\theta_{sk}^{\text{standardized common}}$, the fitted surface averaged over the shared target. They obey the exact identity

$$\theta_{sk}^{\text{standardized common}} - \theta_s^{\text{raw full}} = (\theta_{sk}^{\text{raw support}} - \theta_s^{\text{raw full}}) + (\theta_{sk}^{\text{model own support}} - \theta_{sk}^{\text{raw support}}) + (\theta_{sk}^{\text{standardized common}} - \theta_{sk}^{\text{model own support}}).$$

We label the three terms **support restriction**, **in-support model calibration**, and **modeled target distribution**. This algebraic, model-based decomposition is descriptive; it is neither causal nor a pure composition decomposition.

4.4 Matched four-dimensional plausible values and complex-sample covariance

TIMSS documents that the five longitudinal achievement plausible values were simultaneously imputed from country-specific four-dimensional models spanning mathematics 2023, mathematics 2024, science 2023, and science 2024. Evidence appears in Technical Report Section 6 (printed p. 32 and pp. 36–37) and User Guide Chapter 2 (printed pp. 51–52; Exhibit 2.5) (Fishbein et al., 2025; Fishbein, Taneva, et al., 2026). The terminal digit $j = 1, \dots, 5$ is the plausible-value index. Thus Grade 4-to-5 draw j is the matched tuple $(ASMMAT1j, ASMMAT2j, ASSSCI1j, ASSSCI2j)$; the analogous Grade 8-to-9 tuple is $(BSMMAT1j, BSMMAT2j, BSSSCI1j, BSSSCI2j)$. Except for the separately defined same-wave diagnostic, every inferential primary, standardization, sensitivity, and grade-band extension analysis used these matched tuples. Plausible values were never averaged into person scores. Unequal-index pairs were used only for that explicitly labeled same-wave diagnostic; they were not treated as longitudinal imputations (Mislevy et al., 1992; Wu, 2005).

TIMSS uses two delete-one school or quasi-school replicates per jackknife zone. In a perturbed zone, the retained member receives factor 2 and the omitted member factor 0; all other zones retain factor 1. For estimate vector $\hat{\theta}_j$, sampling covariance was

$$\hat{U}_j = \frac{1}{2} \sum_r (\hat{\theta}_{jr} - \hat{\theta}_j)(\hat{\theta}_{jr} - \hat{\theta}_j)^\top.$$

With $M = 5$, the pooled estimate was $\bar{\theta} = M^{-1} \sum_j \hat{\theta}_j$. The sample between-PV covariance was

219

$$B = \frac{1}{M-1} \sum_{j=1}^M (\hat{\theta}_j - \bar{\theta})(\hat{\theta}_j - \bar{\theta})^\top,$$

220 and total covariance was

221

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B,$$

222 where $\bar{U} = M^{-1} \sum_j U_j$ (Fishbein et al., 2024; Rubin, 1987). Standard errors were square roots of
 223 the diagonal of T ; 95% confidence intervals used $\hat{\theta} \pm 1.96SE$.

224 Within each matched draw, the cross-system/cross-specification stack contained 168 parameters:
 225 nine raw-full quantities, plus the three support/model/standardized quantities for each applicable
 226 specification-system combination. Five specifications contributed nine systems each and the
 227 leave-Korea-out specification contributed eight. Every JRR perturbation rebuilt all affected
 228 percentile bounds, strict-cell boundaries and eligibility, target weights, the perturbed system's
 229 fitted surface, and every standardized estimate affected by the shared target. The primary 18-
 230 parameter covariance (nine raw and nine standardized estimates) is an exact submatrix of this
 231 full stack. It supports covariance-aware standardized-minus-raw and alternative-minus-primary
 232 contrasts without independence assumptions. Pairwise cross-system contrasts were retained only
 233 as an unnumbered exploratory reproducibility diagnostic and were excluded from the publication
 234 supplement to avoid ranking-oriented inference.

235 Canonical artifact replacement stopped if a replicate produced empty support, an invalid target,
 236 rank deficiency, nonconvergence, a nonfinite coefficient or estimate, or a failed covariance gate.
 237 Replicates were never silently dropped. Zones with no BB records for a draw contributed exactly
 238 zero deviation to the risk-set estimand.

239 4.5 Sensitivity and diagnostic analyses

240 We conducted five further classes of analysis.

241 1. **Thresholds and buffers.** We re-estimated transitions at 390, 400, 410, 465, 475, and 485.

242 A 450-to-500 buffer required students to begin below 450 in both domains and finish at

243 or above 500 in both; a low-benchmark buffer used 380-to-420. These define different

244 risk sets and events.

245 2. **Participation and response weighting.** We estimated the primary transition among

246 students indicated as participating in both waves. Separately, among baseline participants,

247 system- and PV-specific quadratic logistic models predicted follow-up participation from

248 baseline mathematics and science. Respondents received stabilized inverse-probability

249 weighting (IPW) for response, with a planned probability floor of .05 (Robins et al., 1994;

250 Seaman & White, 2013). Models, weights, and method contrasts were rebuilt within each

251 JRR replicate. Diagnostics recorded weighted covariate balance before and after

252 weighting, propensity and stabilized-multiplier quantiles and maxima, original and

253 reweighted ESS, and the fraction reaching the floor; support and positivity remain

254 identifying conditions rather than properties guaranteed by model convergence

255 (Westreich & Cole, 2010).

256 3. **Pattern-mixture shifts.** To examine bounded missing-not-at-random (MNAR) scenarios,

257 we subtracted 0, 10, 20, or 40 points from both follow-up-domain PVs for baseline-only

258 students, both baseline-domain PVs for follow-up-only students, and both sets under a

259 combined scenario (Daniels & Hogan, 2008; Little & Rubin, 2019). Delta zero had to

260 reproduce the primary analysis exactly. Interpolated diagnostics were not treated as

261 identified tipping points.

- 262 4. **Exploratory same-wave unequal-index classification-discordance diagnostic.** At each
263 wave, we averaged all 20 ordered unequal-index values of $P(AA_k | BB_j)$, $j \neq k$. Its
264 interval uses JRR sampling variance for the 20-pair arithmetic mean, conditional on the
265 five released PVs; pair dispersion is reported separately and is not a Rubin between-PV
266 component. This diagnostic is neither a longitudinal estimate nor an adjustment for the
267 transition or regression to the mean (Barnett et al., 2005).
- 268 5. **Grade 8-to-9 extension.** We repeated transition, threshold, missingness, and domain-
269 specific analyses in the three available systems as a descriptive cross-grade extension, not
270 an independent replication. Jordan science was interpreted separately because TIMSS
271 documents a change in assessment/instructional organization across waves (Davies et al.,
272 2025).

273 All analyses used base R 4.6.1. Scripts, source manifests, validation outputs, tables, and figures
274 form a sequential reproducible pipeline. No third-party R packages are required.

275 4.6 Generative-AI disclosure

276 During preparation of this manuscript, the author used OpenAI ChatGPT 5.6 Sol model to assist
277 with initial drafting and English-language refinement. The author critically reviewed and revised
278 all AI-assisted material and takes full responsibility for the accuracy, integrity, interpretations,
279 and final content of the manuscript. The AI tool is not credited as an author.

280 5 Results

281 5.1 Data validation and empirical distinction from official marginals

282 The extraction audit covered 24 official source rows, not 24 newly estimated joint statistics.
283 Those rows yielded 108/108 scalar point-estimate comparisons and 102/102 numeric standard-
284 error comparisons within published rounding tolerances; six additional official benchmark
285 standard errors were suppressed and handled as such. Maximum absolute discrepancies were
286 0.4945 TIMSS points for means/growth, 0.4924 percentage points for benchmark percentages,
287 0.0479 points for numeric mean/growth standard errors, and 0.0484 percentage points for
288 numeric benchmark standard errors. There were no missing longitudinal plausible-value cells in
289 the reduced files. These checks validate extraction and weighting, not the substantive joint
290 estimates.

291 Official Grade 4-to-5 mean changes across mathematics and science were 10–42 TIMSS points.
292 At the official 400 benchmark, differences between the separately published rounded at/above
293 shares were 0–5 percentage points for mathematics and 1–9 points for science. Those are
294 differences formed from rounded, domain-specific marginal percentages for the weighted cohort;
295 they are not separately published unrounded change statistics. By contrast, the conditional joint
296 BB-to-AA transition at the same 400 cut score ranged from 14.0% to 47.0%; its denominator is
297 students below 400 in *both* domains at baseline and its numerator is those same students at/above
298 400 in both domains at follow-up. The quantities therefore differ in denominator and meaning
299 and are not competing estimates of one parameter.

At 475, weighted baseline BB shares ranged from 4.1% in the Republic of Korea to approximately 60.1%–60.2% in Kosovo and Jordan. The baseline conditional-risk populations thus differed greatly in prevalence and distance below the two cut scores.

5.2 Four-state movement and the BB-to-AA transition

Grade 4-to-5 BB-to-AA estimates ranged from 5.5% in Kosovo to 34.1% in Slovenia (Figure 1; Table 2). Estimates were 16.6% (95% CI 12.6%–20.5%) in Georgia, 28.2% (24.4%–31.9%) in Italy, 11.9% (9.3%–14.4%) in Jordan, 32.5% (19.9%–45.1%) in the Republic of Korea, 5.5% (4.0%–7.1%) in Kosovo, 18.7% (15.5%–21.9%) in Montenegro, 10.6% (8.2%–13.0%) in North Macedonia, 34.1% (29.7%–38.5%) in Slovenia, and 25.4% (20.2%–30.7%) in Sweden.

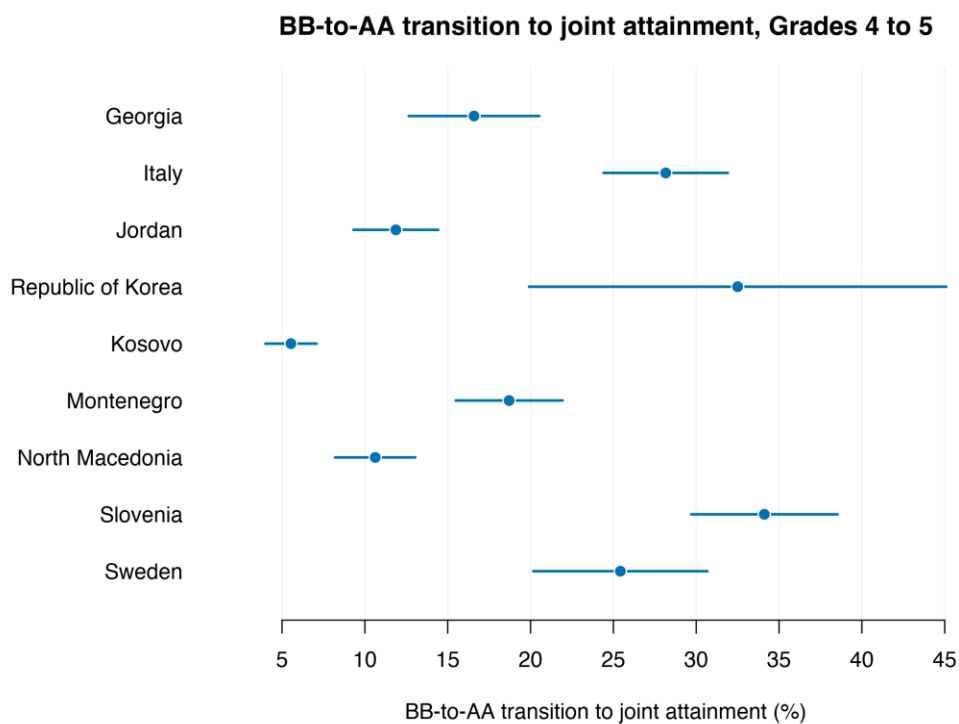


Figure 1: Among Grade 4-to-5 students below both the 475-point mathematics and science thresholds (BB) at baseline, the percentage at or above both (AA) at follow-up. Points combine five matched plausible-value (PV) draws; horizontal bars are jackknife repeated replication (JRR) 95% confidence intervals.

BB persistence ranged from 38.0% in the Republic of Korea and 39.6% in Slovenia to 75.1% in Jordan and 76.7% in Kosovo. Movement from BB into one-domain states (MB or SB) ranged from 13.1% in Jordan to 29.5% in the Republic of Korea. Among students beginning in MB, AA was the largest follow-up destination in all nine systems (39.6%–78.7%). Among those beginning in SB, AA was the largest destination in six systems (34.6%–76.6%); BB was largest in Jordan, while SB persistence was largest in Kosovo and North Macedonia. AA persistence ranged from 71.5% in Jordan to 99.6% in the Republic of Korea, so movement from AA to any other state ranged from 0.4% to 28.5%. The complete matrices and cell-specific uncertainty are provided in Additional file 2 (Table S1); Additional file 1 (Figure S1) visualizes the Grade 4-to-5 point estimates. The probability of reaching AA from *any* baseline deficit ranged from 14.4% in Kosovo to 58.4% in the Republic of Korea. Thus, neither overall AA change nor the focal BB-to-AA transition substitutes for the complete joint matrix.

Among baseline-BB students, mean weakest-domain gains were positive in all nine systems, ranging from 18.7 points in Kosovo to 50.0 points in Slovenia. Positive mean gain can coexist with a low transition probability because students may start far below 475, gain in only one domain, or remain short of one cut score.

5.3 Common-severity estimates and specification sensitivity

Under the primary specification, common-severity standardized estimates ranged from 8.6% in Kosovo to 25.6% in Slovenia, compared with a raw range of 5.5%–34.1% (Figure 2; Table 2). Standardized-minus-raw differences ranged from –9.8 percentage points (95% CI –18.0 to –1.7) in the Republic of Korea to +6.2 points (4.3 to 8.2) in Jordan. These changes describe modeled predictions under the specified equal-system severity target; they do not partition observed differences into causal or pure-composition components.

Observed and common-severity standardized BB-to-AA transition

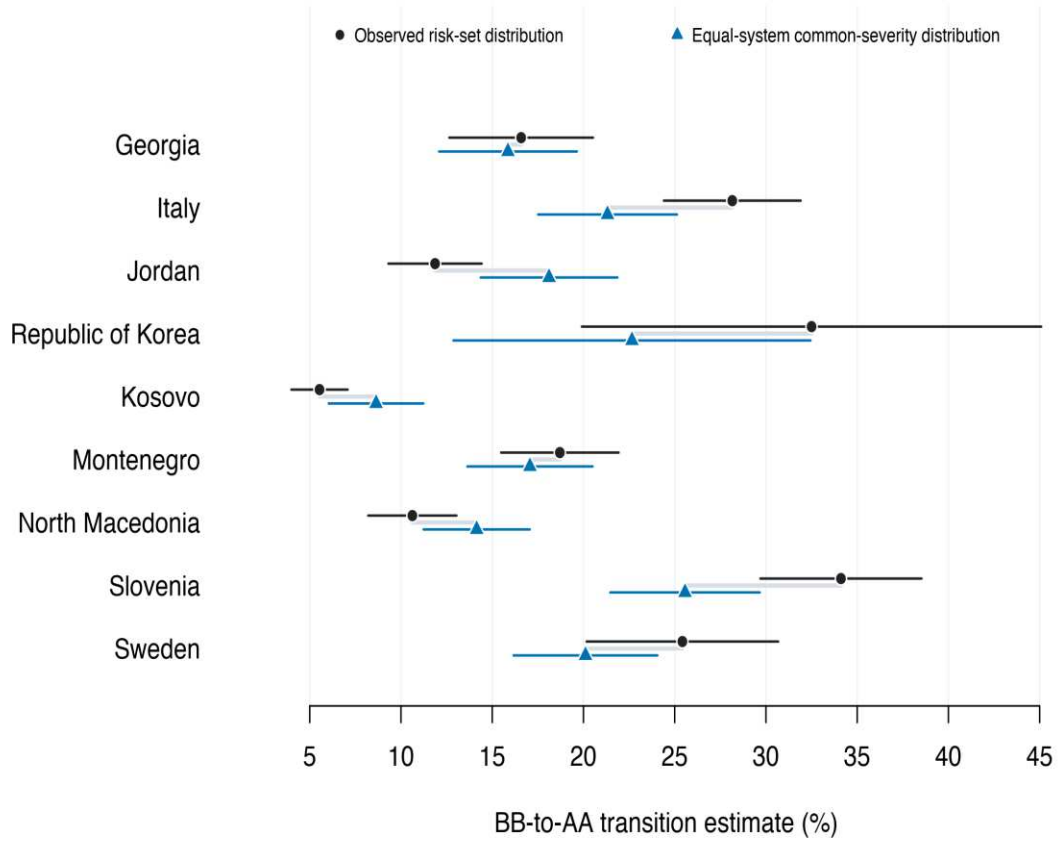


Figure 2: Observed baseline-BB risk-set estimates and modeled estimates integrated over an equal-system common-severity distribution on shared support. Circles and triangles identify the two estimands, connecting lines pair estimates within systems, and horizontal bars are covariance-aware 95% confidence intervals.

332 **Table 2. Raw and common-severity standardized Grade 4-to-5 BB-to-AA transition at 475**

System	Mean BB <i>n</i>	Baseline BB %	Raw BB-to- AA, % (95% CI)	Standardize d BB-to- AA, % (95% CI)	Standardize d – raw, pp (95% CI)	Minimum support ESS
Georgia	1,493.6	34.2	16.6 (12.6– 20.5)	15.9 (12.1– 19.6)	–0.7 (–2.8– 1.4)	896.5
Italy	974.6	21.4	28.2 (24.4– 31.9)	21.3 (17.5– 25.1)	–6.8 (–10.1– –3.6)	642.0
Jordan	4,001.4	60.2	11.9 (9.3– 14.4)	18.1 (14.4– 21.9)	+6.2 (4.3– 8.2)	1,285.9
Republi c of Korea	162.8	4.1	32.5 (19.9– 45.1)	22.7 (12.9– 32.4)	–9.8 (–18.0– –1.7)	88.3
Kosovo	2,701.8	60.1	5.5 (4.0– 7.1)	8.6 (6.0– 11.2)	+3.1 (1.4– 4.8)	1,419.5
Monten egro	1,945.0	42.0	18.7 (15.5– 21.9)	17.1 (13.6– 20.5)	–1.6 (–3.2– –0.1)	1,451.5
North Macedo nia	2,257.2	46.3	10.6 (8.2– 13.0)	14.1 (11.2– 17.1)	+3.5 (1.7– 5.3)	1,327.6
Slovenia	946.6	18.9	34.1 (29.7– 38.5)	25.6 (21.5– 29.7)	–8.5 (–11.7– –5.4)	642.5
Sweden	794.6	15.6	25.4 (20.2– 30.7)	20.1 (16.2– 24.1)	–5.3 (–8.6– –2.1)	418.5

333 *Note.* BB = below both mathematics and science thresholds; ESS = Kish effective sample size;
334 pp = percentage points. Mean BB *n* averages the unweighted risk-set count across five plausible-
335 value draws. Confidence intervals use total matched-PV/JRR covariance.

336 The six-specification artifact contained 53 applicable specification-system rows. The five
337 alternatives yielded 44 covariance-aware alternative-minus-primary contrasts. Point differences
338 ranged from –1.0 to +3.7 percentage points, and none of the individual 95% intervals excluded

zero. This absence of individually distinguishable contrasts does not establish equivalence: standard errors reached 7.9 points, and the widest strict-cell interval extended from -12.3 to $+18.6$ points. The pooled represented-BB target, leave-Korea-out target, and two lower-complexity surfaces were estimated relatively close to the primary, whereas the strict-common-cell specification retained 10–11 of 16 cells across draws and 75.83%–85.89% of target mass before renormalization. Its much wider intervals leave room for nontrivial differences and identify support-rebuilding instability; it remains a sensitivity analysis, not a preferred replacement. All 135 full-sample fits converged at full rank, and the fail-fast pipeline completed all planned replicate refits without dropping a replicate. The 168-parameter covariance had minimum eigenvalue $-4.275395e-18$, above the $-1e-8$ numerical tolerance.

Korea’s evidence must be read as one qualified set: its mean 475 baseline-BB count was 162.8; raw and standardized estimates were 32.5% (19.9%–45.1%) and 22.7% (12.9%–32.4%); minimum support ESS was 88.3; mean target mass within its convex hull was 92.7%; and mean target mass in five-by-five cells with at least five Korea cases was 58.0%. At 400, its mean baseline-BB count was approximately 26 and its standard error was 19.5 percentage points. These denominator and overlap limitations preclude ranking or causal interpretation.

5.4 Threshold, participation, and pattern-mixture analyses

Threshold choice changed both risk-set composition and the transition event (Figure 3). At 400, conditional Grade 4-to-5 BB-to-AA estimates ranged from 14.0% in Kosovo to 47.0% in Slovenia; at 475 they ranged from 5.5% to 34.1%. Most point estimates declined as the threshold increased. Korea’s sparse low-threshold denominators and 19.5-point standard error at 400 show why its curve should not be read as a smooth system trait.

Threshold sensitivity of the BB-to-AA transition

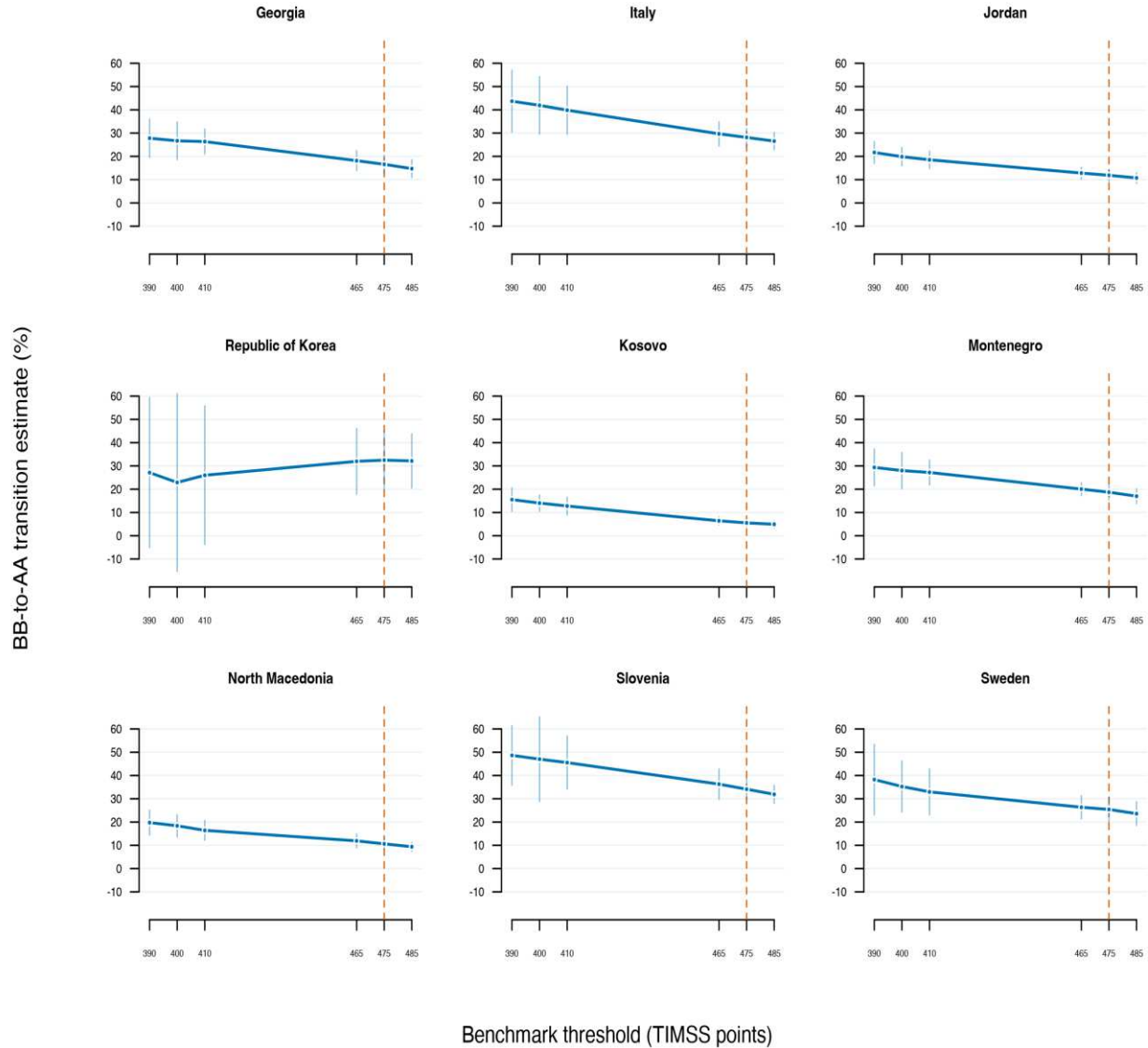


Figure 3: System-specific Grade 4-to-5 BB-to-AA percentages across six planned TIMSS thresholds. Each threshold changes both the baseline-BB denominator and follow-up-AA event; the orange dashed line marks the primary 475-point threshold, and vertical bars are untruncated normal-approximation 95% confidence intervals.

361 The 450-to-500 buffer produced much smaller probabilities, from 0.8% in Kosovo to 8.0% in
 362 Slovenia. These values describe a more demanding event—beginning at least 25 points below
 363 the primary cut in both domains and ending at least 25 points above it in both—and are not
 364 corrected estimates of the primary transition.

365 Both-wave observed and response-IPW results were close to the PV-completed estimates (Figure
366 4; Table 3). Response-IPW point estimates differed from the primary by less than 1 percentage
367 point in eight systems; Sweden differed by -1.3 points. All response models converged, no
368 estimated propensity reached the .05 floor, and effective sample sizes were nearly unchanged.
369 Table S10 reports covariance-aware method-minus-primary contrasts; the accompanying
370 diagnostics report propensity and stabilized-weight-multiplier distributions, covariate balance,
371 ESS, and positivity by system and PV draw. These analyses address selection through observed
372 baseline predictors but cannot rule out selection on unobserved outcomes.

373 Under the combined -40 -point pattern-mixture scenario, BB-to-AA estimates ranged from 5.3%
374 to 32.9%. The largest reduction from delta zero was 2.2 points in Sweden; Korea changed by
375 -1.1 points. No curve reached half its delta-zero estimate or crossed its delta-zero lower 95%
376 confidence bound within the planned range. These are imposed-scenario diagnostics, not
377 evidence that nonresponse is ignorable.

Missing-data and response-weighting sensitivity

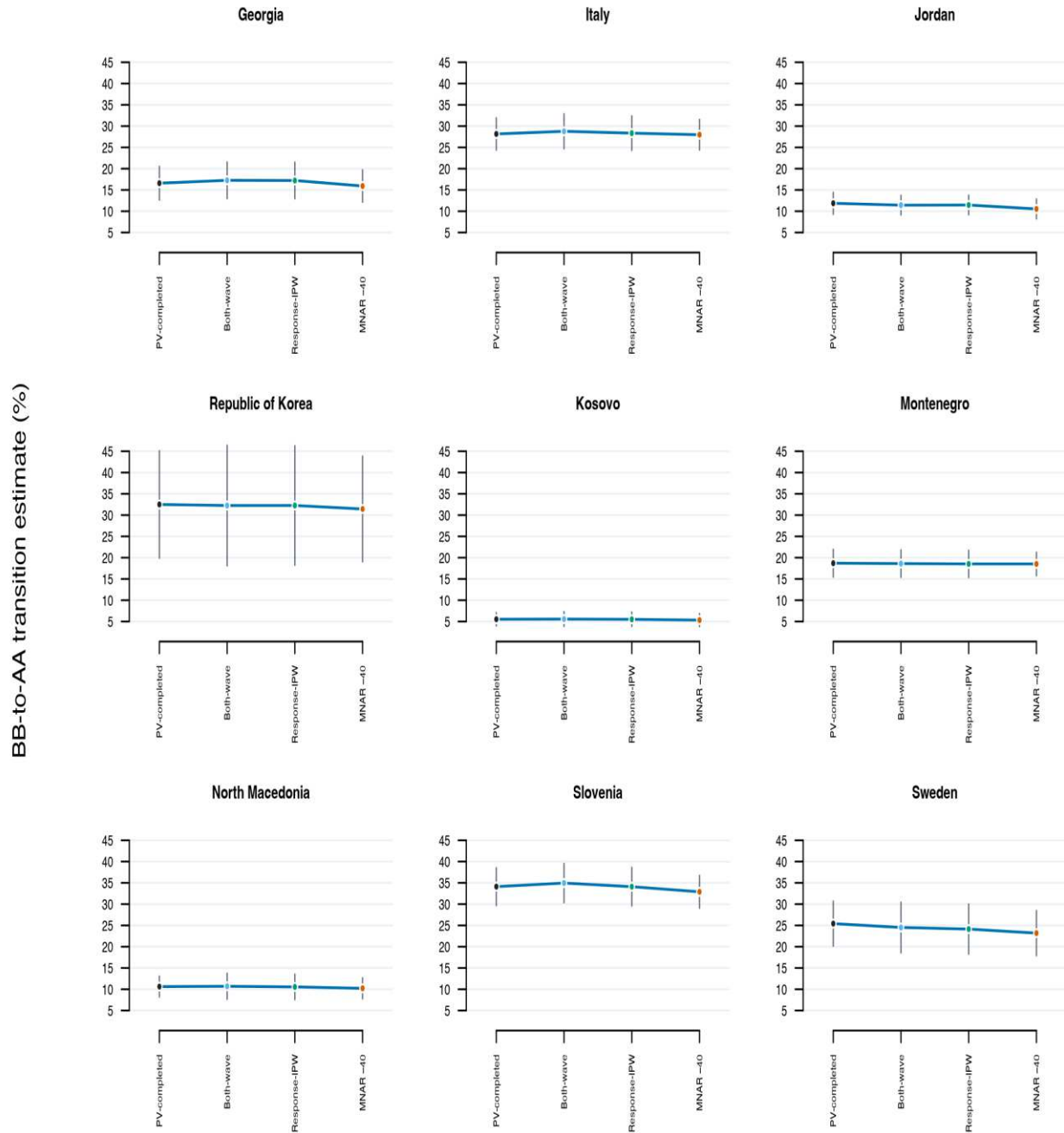


Figure 4: Grade 4-to-5 BB-to-AA percentages under plausible-value (PV) completion, restriction to both-wave participants, inverse-probability weighting (IPW) for observed response, and a combined missing-not-at-random (MNAR) minus-40-point scenario. The methods have different targets or assumptions; vertical bars are 95% confidence intervals.

378 **Table 3. Selected robustness results for the Grade 4-to-5 BB-to-AA transition (percent)**

Syste m	Primary 475	Std. 475	Cut 400	450→50 0 buffer	Both wave	Respons e IPW	MNAR –40	Follow- up discordance
Georg ia	16.6	15.9	26.7	3.9	17.3	17.2	15.9	5.9
Italy	28.2	21.3	41.9	6.0	28.8	28.3	28.0	7.6
Jorda n	11.9	18.1	19.9	4.5	11.4	11.5	10.5	3.5
Repu blic of Korea	32.5	22.7	22.9	4.9	32.3	32.3	31.4	9.1
Koso vo	5.5	8.6	14.0	0.8	5.6	5.5	5.3	1.8
Mont enegr o	18.7	17.1	28.0	4.9	18.6	18.5	18.5	5.7
North Mace donia	10.6	14.1	18.4	3.0	10.7	10.6	10.2	2.9
Slove nia	34.1	25.6	47.0	8.0	34.9	34.1	32.9	11.1
Swed en	25.4	20.1	35.3	5.4	24.5	24.1	23.2	11.1

379 *Note.* Std. 475 is common-severity standardized at 475. Cut 400 and the 450→500 buffer define
380 different denominators and events. Both wave denotes observed both-wave participants;
381 response IPW denotes the inverse-probability-weighted response analysis. MNAR –40 is an
382 imposed missing-not-at-random score-shift scenario. Follow-up discordance is the exploratory
383 follow-up same-wave unequal-index classification-discordance diagnostic; its interval is JRR
384 sampling-only conditional on five released plausible values, not an adjusted longitudinal estimate.
385 Complete standard errors and intervals are in Additional files 1 and 2.

5.5 Exploratory same-wave classification-discordance diagnostic

Same-wave unequal-index BB-to-AA classification-discordance means were below the matched-PV longitudinal transition in all nine systems. At follow-up, values ranged from 1.8% in Kosovo to 11.1% in Sweden and represented approximately 27%–44% of the corresponding longitudinal point estimates. Baseline values ranged from 1.4% to 6.8%. Plausible-value classification variation is therefore not negligible, but this exploratory diagnostic cannot be subtracted as a pure-error correction: plausible values are posterior population draws rather than repeat forms, and its intervals contain JRR sampling variance only, conditional on the five released PVs.

5.6 Descriptive Grade 8-to-9 extension

In this separate cross-grade extension, BB-to-AA estimates were 5.4% in Jordan (95% CI 3.8%–7.0%), 6.8% in the Republic of Korea (3.0%–10.6%), and 8.1% in Sweden (5.9%–10.3%). Continuous weakest-domain changes were 22.7 points in Jordan, 7.2 points in Korea, and 3.8 points in Sweden; the latter two intervals included zero. Domain-specific below-to-at/above transitions were larger: mathematics estimates were 12.2%, 30.8%, and 25.4%, and science estimates were 13.8%, 22.1%, and 18.6%, respectively. This extension is not an independent replication of the nine-system Grade 4-to-5 estimand. Jordan’s science result remains a separate descriptive result because of the documented organizational change (Davies et al., 2025).

6 Discussion

6.1 Principal findings

The BB-to-AA transition answers a conditional cross-domain question that official domain-specific marginals do not identify. Official Grade 4-to-5 mean changes of 10–42 points and differences between separately published rounded 400-point at/above shares of 0–5 percentage

408 points in mathematics and 1–9 in science describe entire weighted cohorts one domain at a time.
409 Conditional joint transitions at 400 were 14.0%–47.0% among students below both cuts at
410 baseline. Complete four-state matrices showed additional heterogeneity: BB-to-one-domain
411 movement ranged from 13.1% to 29.5%, while AA persistence ranged from 71.5% to 99.6%.
412 The matrix therefore distinguishes persistence, one-domain movement, joint attainment, and
413 movement out of AA that no marginal pair or focal cell can recover.

414 At 475, raw BB-to-AA estimates ranged from 5.5% to 34.1%; estimates integrated over the
415 primary equal-system severity target ranged from 8.6% to 25.6%. The estimates changed under
416 common severity, but that change should not be described as a clean separation of composition
417 and performance. Support restriction, in-support model calibration, and modeled target
418 distribution are algebraic, specification-dependent components. They describe how observed and
419 fitted quantities connect under the model; they do not identify causal mechanisms or a pure
420 composition effect.

421 The six-specification analysis adds two qualifications. First, no one of the 44 alternative-minus-
422 primary intervals excluded zero at the 95% level; observed point differences were –1.0 to +3.7
423 percentage points. This means that the analysis did not distinguish an alternative from the
424 primary at the reported precision, not that the specifications were equivalent. Strict-cell standard
425 errors reached 7.9 points, leaving those intervals compatible with nontrivial differences. Second,
426 strict common cells retained only 10–11 of 16 cells and produced wide intervals despite meeting
427 the frozen n and ESS rules. This is evidence of support instability, not confirmation that the
428 stricter point estimates are more credible.

Korea requires especially cautious interpretation. Its risk set averaged 162.8 students, its raw and standardized intervals were wide, minimum support ESS was 88.3, convex-hull coverage was 92.7%, and five-case-cell coverage was 58.0%. Its threshold-400 denominator averaged approximately 26 and its standard error was 19.5 points. Reporting these quantities together is more informative than placing Korea in a rank order.

Missingness comparisons were comparatively stable within their modeled ranges, while threshold and buffer changes materially changed the event. This is expected: shifting a threshold changes both who enters the denominator and what counts as joint attainment. Same-wave crossing also remained nonzero, so categorical transitions should be accompanied by continuous changes and measurement diagnostics rather than treated as error-free person classifications.

6.2 Contribution to large-scale-assessment reporting

The contribution is the integrated estimand-and-inference workflow, not a new estimator. The four-state matrix aligns the analysis with a multivariate question; common-severity standardization makes the baseline target explicit; matched four-dimensional PV tuples preserve the documented imputation structure; and the 168-parameter replicate/PV covariance carries shared-support and shared-target uncertainty into within- and cross-specification contrasts. Holding the common target fixed across replicates would omit covariance induced when one system's perturbation changes the integration distribution used for multiple estimates.

This workflow complements rather than replaces official reporting. Mean growth summarizes average scale change; marginal benchmark shares summarize each domain; the four-state matrix and BB-to-AA transition answer conditional cross-domain questions. A concise reporting set

could combine those marginals with complete joint matrices, weakest-domain change, and overlap-qualified standardized estimates.

6.3 Strengths

The analysis uses official student weights, five matched four-dimensional PV tuples, and the documented two-replicate-per-zone JRR factor. The audit evaluated 24 source rows and reproduced 108/108 scalar points and 102/102 numeric standard errors while handling six suppressed values. The raw transition and standardization pipelines reproduced primary points and standard errors exactly. Across six specifications, all 135 full-sample models converged at full rank, no standardization replicate failed, all 13 sensitivity checks and all 11 standardization-producer checks passed, and the 168-parameter covariance met its $-1e-8$ eigenvalue tolerance. The generated tables, figures, manuscript fragments, and diagnostics passed all 17 publication-output gates. These are computational and extraction checks; they do not convert a descriptive comparison into a causal one.

6.4 Limitations

Several limitations constrain interpretation.

First, results generalize only to each TIMSS target population represented by the participating longitudinal sample and its weights. They do not estimate a global average or support inference to nonparticipating systems.

Second, benchmark states are constructed from plausible values intended for population inference, not individual diagnosis. Even with correct pooling, a student-level state should not be treated as an observed clinical-style status.

471 Third, 475 was selected before the BB-to-AA estimates were calculated because it yielded a
472 more feasible baseline-BB denominator in all nine systems; the study was not externally
473 preregistered. The transition remains threshold-dependent; Korea's low-threshold estimates were
474 especially sparse, and buffered transitions answer different questions.

475 Fourth, direct standardization is model-based. The quadratic logit surface may miss local
476 nonlinearities. A marginal rectangle, convex-hull diagnostic, and grid checks do not eliminate
477 extrapolation, and the strict-cell alternative revealed instability when joint support was tightened.

478 Fifth, the three decomposition components are not independently identified mechanisms. Their
479 values depend on support, target, and response-surface choices and do not isolate distributional
480 composition from conditional performance.

481 Sixth, participation analyses do not identify missing outcomes. Response-IPW assumes
482 conditional exchangeability under observed baseline scores, and pattern-mixture shifts impose
483 rather than estimate departures from the PV-completed representation.

484 Seventh, the exploratory same-wave unequal-index classification-discordance analysis is not a
485 repeat administration or classical measurement-error estimate. Its intervals are JRR sampling-
486 only, conditional on the five released PVs; it cannot correct the longitudinal transition, supply a
487 Rubin between-PV component, or isolate regression to the mean.

488 Eighth, the descriptive Grade 8-to-9 extension includes only three systems, is not an independent
489 replication of the nine-system Grade 4-to-5 estimand, and includes a change in Jordan's science
490 organizational context across waves (Davier et al., 2025).

The analysis also does not model schools, classrooms, curricula, or student-background predictors. Future work could examine predictors while preserving matched-PV and replicate uncertainty, but this study does not attribute transitions to mechanisms.

7 Conclusions

TIMSS 2023 Longitudinal permits a conditional cross-domain question that separate mathematics and science marginals cannot answer: at the primary 475-point threshold, what proportion of the baseline BB risk set was at or above both benchmarks at follow-up? Grade 4-to-5 BB-to-AA point estimates varied across nine systems, and system-specific point estimates changed when fitted surfaces were integrated over the specified shared equal-system baseline-severity target. None of the 44 covariance-aware alternative-minus-primary 95% confidence intervals excluded zero; this does not establish equivalence, and the strict-common-cell intervals remained compatible with nontrivial differences. The official marginal comparison, Korea's sparse denominator and weaker overlap, and the descriptive Grade 8-to-9 extension all reinforce the need to report denominators, targets, and uncertainty.

Longitudinal large-scale-assessment reporting should pair marginal growth and benchmark shares with joint transition matrices when questions concern coordinated attainment. Common-severity estimates can show sensitivity to a modeled target, but they do not cleanly separate composition and performance and are not causal. Used within these descriptive limits, the BB-to-AA transition adds a population-level dimension that marginal benchmarks alone cannot provide.

510

511 8 Abbreviations

512 AA: at or above both mathematics and science thresholds; BB: below both thresholds; CI:
513 confidence interval; ESS: effective sample size; IPW: inverse-probability weighting; JRR:
514 jackknife repeated replication; MB: mathematics below only; MNAR: missing not at random; pp:
515 percentage points; PV: plausible value; SB: science below only; TIMSS: Trends in International
516 Mathematics and Science Study.

517 9 Declarations

518 9.1 Ethics approval and consent to participate

519 Not applicable to this secondary analysis. The study used publicly available, deidentified TIMSS
520 2023 Longitudinal public-use data and involved no new participant recruitment, contact,
521 intervention, or access to direct identifiers. The original data collection, consent, and governance
522 procedures were administered by the responsible international and national study centers under
523 their documented protocols (Fishbein et al., 2025; Fishbein, Taneva, et al., 2026).

524 9.2 Consent for publication

525 Not applicable; the manuscript contains no identifiable individual information.

526 9.3 Availability of data and materials

527 The raw TIMSS 2023 Longitudinal international database, codebooks, user guide, technical
528 report, and official result workbooks are publicly available from the TIMSS 2023 website. The
529 reproducibility package includes acquisition code and a source-checksum manifest but does not
530 redistribute raw data. Derived analysis files can be regenerated with Rscript analysis/run_all.R;

its acquisition stage downloads and validates the official archives and therefore requires internet access, while subsequent stages can use the validated local files. Complete machine-readable result tables accompany the manuscript in Additional file 2.

9.4 Code availability

All base-R analysis scripts, the standard-library XLSX extractor and package builder, configuration, validation diagnostics, Quarto sources, and a build recipe are included in the anonymous review package in Additional file 3. A Zenodo DOI has been reserved for an unpublished draft record, but the associated record may contain identifying metadata; the DOI is therefore intentionally withheld from this double-anonymous review version. It will be disclosed only after anonymity is lifted and the repository record is published.

9.5 Competing interests

The author declares that there are no competing interests.

9.6 Funding

No funding was received for this research.

9.7 Authors' contributions

The sole author was responsible for Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, and Writing – review & editing. The author reviewed and approved the final manuscript.

550 9.8 Acknowledgements

551 None.

552 10 Additional files

553 • **Additional file 1:** Additional_file_1_supplement.docx (DOCX), *Supplementary methods,*
554 *results, and reproducibility record.* Contains detailed estimands and covariance
555 derivations, selected generated summaries, supplemental figure callouts, diagnostics,
556 sensitivities, and the artifact inventory; complete machine-readable tables are supplied in
557 Additional file 2.

558 • **Additional file 2:** Additional_file_2_machine_readable_tables.zip (ZIP), *Machine-*
559 *readable publication tables.* Contains all 19 publication CSV tables (Tables 1–3 and S1–
560 S16), the ordered table manifest, and an internal SHA-256 manifest.

561 • **Additional file 3:** Additional_file_3_anonymous_reproducibility.zip (ZIP), *Anonymous*
562 *reproducibility package.* Contains analysis code, anonymous manuscript and supplement
563 sources, the package builder and build recipe, nonidentifying diagnostics, publication
564 tables and figures, the quarantined unnumbered exploratory pairwise-contrast file,
565 runtime information, and official-source checksums; raw and derived student data,
566 confidential submission files, and author metadata are excluded.

567

568

569 11 References

570 Barnett, A. G., Pols, J. C. van der, & Dobson, A. J. (2005). Regression to the mean: What it is
571 and how to deal with it. *International Journal of Epidemiology*, 34(1), 215–220.

572 <https://doi.org/10.1093/ije/dyh299>

573 Berger, N., Mackenzie, E., & Holmes, K. (2020). Positive attitudes towards mathematics and
574 science are mutually beneficial for student achievement: A latent profile analysis of TIMSS 2015.
575 *The Australian Educational Researcher*, 47(3), 409–444. [https://doi.org/10.1007/s13384-020-](https://doi.org/10.1007/s13384-020-00379-8)
576 00379-8

577 Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for*
578 *bayesian modeling and sensitivity analysis*. Chapman & Hall/CRC.

579 <https://doi.org/10.1201/9781420011180>

580 Davier, M. von, Kennedy, A. M., Reynolds, K. A., Fishbein, B., Khorramdel, L., Aldrich, C.,
581 Bookbinder, A., Bezirhan, U., & Yin, L. (2024). *TIMSS 2023 international results in*
582 *mathematics and science*. TIMSS & PIRLS International Study Center, Boston College.

583 <https://doi.org/10.6017/lse.tpisc.timss.rs6460>

584 Davier, M. von, Kennedy, A. M., Reynolds, K. A., Gonzalez, E., & Khorramdel, L. (2025).
585 *TIMSS 2023 longitudinal international results in mathematics and science*. TIMSS & PIRLS
586 International Study Center, Boston College. <https://doi.org/10.6017/lse.tpisc.timss.dn2487>

587 Fishbein, B., Bezirhan, U., & Davier, M. von. (2026). International measures of student growth
588 in mathematics and science achievement: The TIMSS 2023 longitudinal international database.
589 *Chinese/English Journal of Educational Measurement and Evaluation*, 7(1), 5.

590 <https://doi.org/10.59863/ykdj2701>

591 Fishbein, B., Gonzalez, E., & Davier, M. von (Eds.). (2025). *TIMSS 2023 longitudinal technical*
592 *report: Addendum to the TIMSS 2023 technical report*. TIMSS & PIRLS International Study
593 Center, Boston College. <https://doi.org/10.6017/lse.tpisc.timss.de4812>

594 Fishbein, B., Siegel, P., Atasever, U., & Cooney, D. (2024). Standard error estimation in TIMSS.
595 In *TIMSS 2023 technical report* (Chapter 13). TIMSS & PIRLS International Study Center,
596 Boston College. <https://doi.org/10.6017/lse.tpisc.timss.rs6950>

597 Fishbein, B., Taneva, M., & Kowolik, K. (2026). *TIMSS 2023 longitudinal user guide for the*
598 *international database*. TIMSS & PIRLS International Study Center, Boston College.
599 https://timss2023.org/wp-content/uploads/data/T23L_UG-International-Database.pdf

600 Huang, F. L. (2024). Using plausible values when fitting multilevel models with large-scale
601 assessment data using R. *Large-Scale Assessments in Education*, 12(1), 7.
602 <https://doi.org/10.1186/s40536-024-00192-0>

603 Jewsbury, P. A., Jia, Y., & Gonzalez, E. J. (2024). Considerations for the use of plausible values
604 in large-scale assessments. *Large-Scale Assessments in Education*, 12(1), 24.
605 <https://doi.org/10.1186/s40536-024-00213-y>

606 Kaur, T., McLoughlin, E., & Grimes, P. (2022). Mathematics and science across the transition
607 from primary to secondary school: A systematic literature review. *International Journal of*
608 *STEM Education*, 9(1), 13. <https://doi.org/10.1186/s40594-022-00328-0>

609 Keiding, N., & Clayton, D. (2014). Standardization and control for confounding in observational
610 studies: A historical perspective. *Statistical Science*, 29(4), 529–558. [https://doi.org/10.1214/13-](https://doi.org/10.1214/13-STS453)
611 STS453

612 Lane, P. W., & Nelder, J. A. (1982). Analysis of covariance and standardization as instances of
 613 prediction. *Biometrics*, 38(3), 613–621. <https://doi.org/10.2307/2530043>

614 Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John
 615 Wiley & Sons. <https://doi.org/10.1002/9781119482260>

616 Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population
 617 characteristics from sparse matrix samples of item responses. *Journal of Educational*
 618 *Measurement*, 29(2), 133–161. <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>

619 Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when
 620 some regressors are not always observed. *Journal of the American Statistical Association*,
 621 89(427), 846–866. <https://doi.org/10.1080/01621459.1994.10476818>

622 Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
 623 <https://doi.org/10.1002/9780470316696>

624 Rutkowski, L., Gonzalez, E., Joncas, M., & Davier, M. von. (2010). International large-scale
 625 assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2),
 626 142–151. <https://doi.org/10.3102/0013189X10363170>

627 Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with
 628 missing data. *Statistical Methods in Medical Research*, 22(3), 278–295.
 629 <https://doi.org/10.1177/0962280210395740>

630 Westreich, D., & Cole, S. R. (2010). Invited commentary: Positivity in practice. *American*
 631 *Journal of Epidemiology*, 171(6), 674–677. <https://doi.org/10.1093/aje/kwp436>

- 632 Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational*
633 *Evaluation*, 31(2–3), 114–128. <https://doi.org/10.1016/j.stueduc.2005.05.005>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1supplement.docx](#)
- [Additionalfile2machinereadabletables.zip](#)
- [Additionalfile3anonymousreproducibility.zip](#)