

Online Appendix — Making School-Set Marks Comparable: Test Difficulty and Hidden Learning Inequality

Peter Courtney

October 22, 2026

Online Appendix

This Online Appendix provides supplementary tables and results referenced in the main text. All tables are reproduced from estimation output; see [Appendix E](#) for replication details.

Appendix A: Attrition Robustness

The main estimation sample requires successful linkage of Grade 7 students to Grade 8 classes. The tables below assess whether attrition at the linkage stage introduces systematic bias into the estimated calibrated learning levels.

Attrition analysis: linked vs joinable students

Balance in student-level attainment is fragile. A particular concern is non-progressors, students who do not enter Grade 8 the following year and so cannot be linked. In 2022, the cohort for which the loss is broken down, 3.2 per cent of Grade 7 students were in Grade 7 again the next year, 9.7 per cent were in Grade 8 but did not link, and 9.3 per cent were in neither grade in the six provinces. Splitting the unlinked students at the cohort's 15th-percentile into likely non-progressors below that threshold and likely progressors above that threshold, linked students sit 2.44 marks below that joinable group. The sign of the gap is an artefact of the cut-off, not a fact about selection. Figure 1 plots it as the cut-off moves from the whole attrited pool to the 30th percentile: linked students sit 5.5 marks above the untruncated pool, the gap crosses zero close to the tenth percentile, where the two groups differ by 0.45 marks, and reaches 5.1 marks below by the thirtieth, because each higher cut-off drops more low-scoring non-progressors from the comparison group and lifts its mean.

Table 1: Attrition analysis: linked students vs. joinable non-linked students.

Characteristic	Linked (In)	Joinable Proxy (Out)	Non- Progressor Proxy (Out)	Difference (In vs Joinable)	Cohen's <i>d</i>	<i>p</i> -value
Mean Grade 7 mark	54.21	56.64	30.30	-2.44	-0.175*	<0.001
% Female	50.33	49.61	30.05	0.72	0.014	<0.001
% African/Black	93.62	93.44	94.42	0.18	0.007	<0.01
% Home language African	90.76	88.33	91.74	2.43	0.083	<0.001
% Over-age	30.82	33.51	62.01	-2.69	-0.058	<0.001
% Quintile 1 schools	25.43	23.57	28.28	1.86	0.043	<0.001

(continued)

Characteristic	Linked (In)	Joinable Proxy (Out)	Non-Progressor Proxy (Out)	Difference (In vs Joinable)	Cohen's d	p -value
% Quintile 5 schools	10.28	12.00	6.74	-1.72	-0.056	<0.001

Linked (In): Grade 7 students matched to a Grade 8 Term 2 record: the analysis sample. **Joinable Proxy (Out):** Unlinked students above the cohort-specific 15th percentile baseline mark, likely progressors the algorithm failed to match; this is the comparison group for the attrition test. **Non-Progressor Proxy (Out):** Unlinked students at or below the 15th percentile, likely repeaters or dropouts, excluded from the comparison and reported for context only. The “Difference (In vs Joinable)” column reports raw differences in means in the units of the row variable (mark points for continuous variables, percentage-point differences for indicator variables). Cohen's d uses the pooled within-group standard deviation of the Linked and Joinable Proxy groups; values with $|d| > 0.10$ are flagged with an asterisk (*). The p -value is from a two-sample t -test. Sample: National (6 provinces), pooled 2018 and 2022 cohorts.

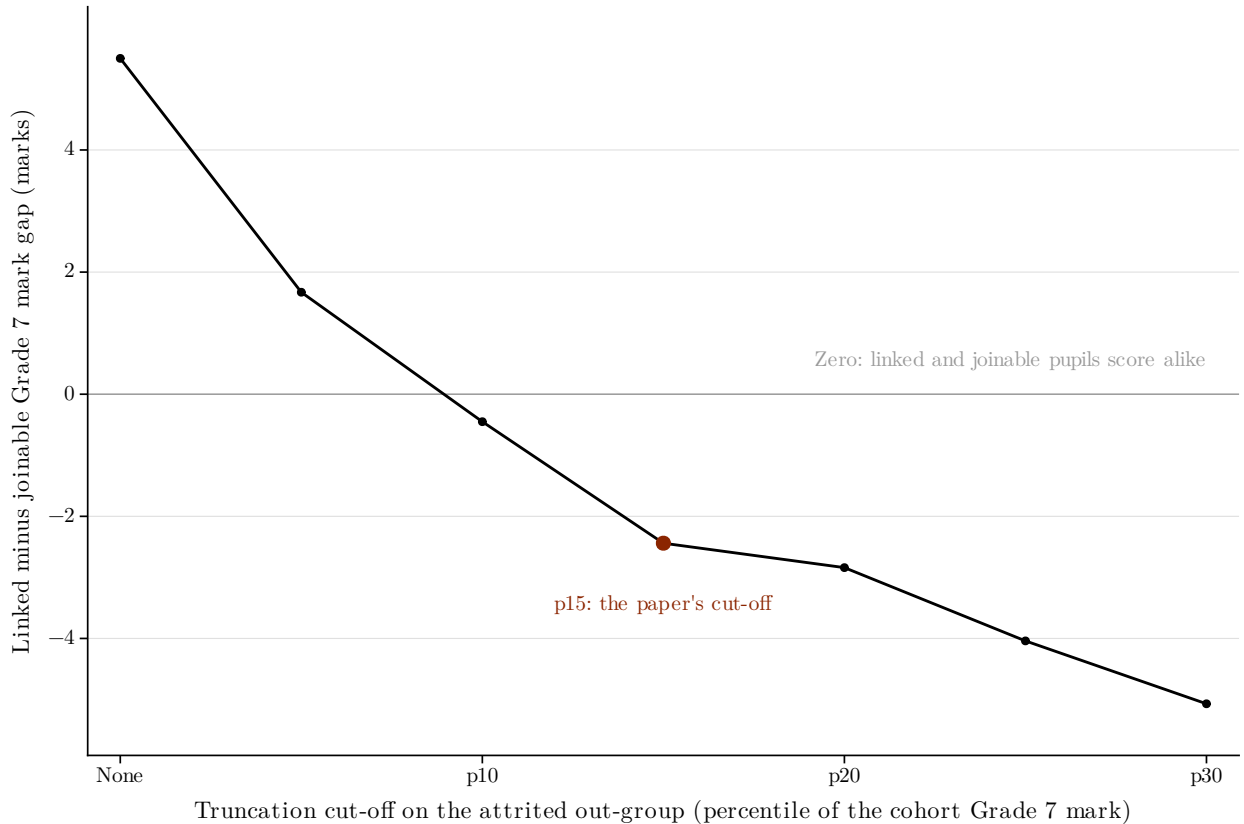


Figure 1: **Whether linked students score above or below the joinable out-group depends on where that out-group is truncated.** *Linked-minus-joinable Grade-7 mark gap as the truncation cut-off moves from none (the whole attrited pool) to the 30th percentile of the cohort baseline. A higher cut-off drops more low-scoring non-progressors from the out-group and lifts its mean, so the gap falls from positive (linked score higher) through zero near the tenth percentile to negative. The paper's 15th-percentile cut-off is highlighted. The cut-off-free inverse-linkage-probability check does not depend on this choice. Sample: National (6 provinces), pooled 2018 and 2022 cohorts.*

Combined school linkage rates

Linkage rates by school type and cohort year. Combined schools (offering both primary and secondary phases) retain a higher fraction of students internally, yielding higher linkage rates.

Table 2: Linkage rates by school type and cohort.

Cohort	School type	N students	N linked	Linkage rate	Mean Gr7 mark
2018	Combined School	67,886	55,453	0.82	53.94
2018	Independent School	9,000	6,477	0.72	57.27
2018	Standalone Public School	631,893	475,691	0.75	58.46
2022	Combined School	75,680	64,169	0.85	55.09
2022	Independent School	11,818	8,828	0.75	61.08
2022	Standalone Public School	717,328	593,806	0.83	59.00

Sample: National (6 provinces), pooled 2018 and 2022 cohorts.

Attrition balance (excluding combined schools)

Balance between linked and unlinked students on observable characteristics, restricting to standalone public schools. Differences are statistically significant but substantively small in standard-deviation units, consistent with mild positive selection on prior marks among linked students.

Table 3: Attrition balance excluding combined schools.

Cohort	Characteristic	In (Linked)	Out (Joinable Proxy)	Difference	Diff (SD)	p-value
2018	Mean Grade 7 mark	59.44	62.45	-3.01	-0.226	<0.001
2018	% Quintile 1 schools	25.13	25.18	-0.06	-0.001	0.699
2018	% Quintile 5 schools	11.58	9.85	1.73	0.055	<0.001
2022	Mean Grade 7 mark	59.85	63.47	-3.61	-0.258	<0.001
2022	% Quintile 1 schools	25.53	21.42	4.11	0.095	<0.001
2022	% Quintile 5 schools	10.81	15.92	-5.10	-0.160	<0.001
Pooled	Mean Grade 7 mark	59.67	62.89	-3.22	-0.235	<0.001
Pooled	% Quintile 1 schools	25.35	23.56	1.80	0.041	<0.001
Pooled	% Quintile 5 schools	11.15	12.47	-1.32	-0.042	<0.001

Sample: National (6 provinces), pooled 2018 and 2022 cohorts.

School-level linkage rate correlates

Correlation between school-level linkage rates and observable school characteristics. The negligible correlation with school quintile indicates that attrition is not differentially concentrated in poorer or wealthier schools.

Table 4: Correlates of school-level linkage rates.

Characteristic	Correlation with linkage rate	Note
Mean Grade 7 mark	0.101	Weak. The test that binds is whether linkage predicts the school effect, reported below.
School Quintile (1=lowest, 5=highest)	-0.031	Negligible; selection is not differential by quintile.

Sample: National (6 provinces), pooled 2018 and 2022 cohorts.

IPW sensitivity check

Inverse-probability-weighted (IPW) re-estimation of the school wedges. The IPW weights are constructed from observed school-level linkage rates; the weight for school s is $1/\text{linkage_rate}_s$. The negligible mean shift under IPW indicates that selection into the linked sample does not materially affect the school-level estimates. See §2.4 of the main text for discussion.

Table 5: IPW sensitivity: the school wedges vs. linkage rates.

Metric	Value	Interpretation
N school-years matched to linkage rates	19,883	Standalone public schools x cohort-year (≥ 5 valid Gr7 students)
Correlation: school effect vs. linkage rate	0.021	Negligible correlation
R-squared: school effect ~ linkage rate	0.000	$< 1\%$ variance explained by linkage rate
Slope: school effect ~ linkage rate (bias proxy)	0.138	Uninformative on its own; see the note
Mean school effect (unweighted)	0.22	Unweighted mean school effect (marks)
Mean school effect (IPW-weighted)	0.22	IPW-weighted mean school effect (marks)
Difference (IPW - unweighted)	-0.005	Negligible – weighting does not shift estimates

The slope of the school effect on the linkage rate is reported for completeness and is uninformative on its own: linkage rates vary across a narrow range, so a slope of this size moves the school effect by a negligible amount over the observed range, which is why the correlation is 0.021 and the R-squared rounds to zero. The correlation pools the twelve province-cohort fits, each of which measures its wedges from its own zero, so it partly compares fits rather than schools; within fits it is -0.034 , and the IPW shift is 0.004 SD (§2.4 of the main text). Inverse-probability weighting upweights schools with lower linkage rates; the IPW-weighted mean is reported against the unweighted baseline. Sample: National (6 provinces), standalone public schools with ≥ 5 valid Grade 7 students, pooled 2018 and 2022 cohorts.

Connected-set drop-out

Schools and students removed by the within-cohort connected-set restriction, summarised by observable partition. Dropped schools are systematically smaller than retained schools but do not differ detectably in poverty quintile, urban/rural status, or no-fee status. See §2.4 of the main text for discussion. I do not compute an Oster (2019) bound: with observed selection into the connected set this small, the bound it would place on unobserved selection is uninformative.

Table 6: Connected-set drop-out: schools and students removed by the within-cohort connected-set restriction, summarised by observable partition (pooled 2018 + 2022).

Partition	Test	Schools dropped		Students dropped		p-value
		n	(%)	n	(%)	
Size (continuous)	Welch t-test	13	0.45	252	0.10	8.6e-12
Poverty quintile	Chi-square	1	0.04	—	—	0.12
Urban / rural	Chi-square	6	0.21	—	—	0.96
No-fee status	Chi-square	6	0.21	—	—	0.074
District	Chi-square	6	0.21	—	—	0.57
Size decile	Chi-square	13	0.45	—	—	2.5e-13

Sample: Gauteng core (Gauteng Mathematics), pooled 2018 and 2022 cohorts.

A.6: Heterogeneous treatment effects under selective attrition

The selective-attrition Monte Carlo of §5.1 of the main text (Table 6 of the main text) is built on a linear-additive data-generating process: the per-student return to ability does not vary across schools. But *heterogeneous* treatment effects (better schools producing larger gains for higher-ability students) could in principle generate bias under selective attrition that the linear-additive simulation cannot detect. This appendix subsection extends that simulation to quantify the bias as a function of interaction strength.

Extended DGP. The seven attrition scenarios of Table 6 of the main text are run unchanged. The endline equation is augmented with a single interaction term:

$$\text{endline}_i = \beta \cdot \text{baseline}_i + \tau_j + \gamma_c + \delta_{\text{het}} \cdot \text{ability}_i \cdot \tilde{\tau}_j + \varepsilon_i,$$

where $\tilde{\tau}_j = \tau_j / \sigma_\tau$ is standardised teaching quality and δ_{het} is a standardised slope perturbation: at a school one SD above the mean in quality, the per-mark-of-ability return to learning is augmented by δ_{het} marks. At (+1 SD ability, +1 SD school) the extra outcome is $\delta_{\text{het}} \cdot \sigma_a \approx 12 \cdot \delta_{\text{het}}$ marks. $\delta_{\text{het}} = 0$ nests the DGP of §5.1 of the main text exactly. We sweep δ_{het} over the grid $\{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.50\}$, crossed with the seven scenarios of §5.1 of the main text, giving a 7×7 grid of 300 simulations per cell.

Headline outcome. Because the school FE is identified up to an additive constant, the per-replication mean bias is mechanically zero. The interpretable headline for the heterogeneity question is the $Q_5 - Q_1$ bias spread: the bias of the school FE in the top quintile of true tau minus the bias in the bottom quintile, in marks. Under linear-additive selective attrition this spread should be ~ 0 ; under heterogeneous-effect selective attrition it grows in δ_{het} in proportion to how much the within-school ability distribution is shifted by selection.

Verification. Because this is a stand-alone Monte Carlo extension, the code includes an automated cross-check: at $\delta_{\text{het}} = 0$ the extension should reproduce the results of §5.1 of the main text exactly. It does: across all seven scenarios the maximum absolute discrepancy in the linked-sample rank correlation is 0.0001 and in the linked-sample RMSE is 0.0022 marks. The full 7×7 cell summary and a self-contained verification report confirm this reproduction.

Findings. Under random attrition the slope of the $Q_5 - Q_1$ bias spread on δ_{het} is -0.060 marks per unit (essentially zero, as the linear-additive analysis predicts: random selection induces no systematic shift in the ability distribution by school quality). Under calibrated attrition (the empirically realistic scenario matching the observed correlation between linkage rates and school quality), the slope is 1.20 marks per unit. Under the extreme scenario (joint strong selection on both ability and school quality) the slope is 2.93 marks per unit.

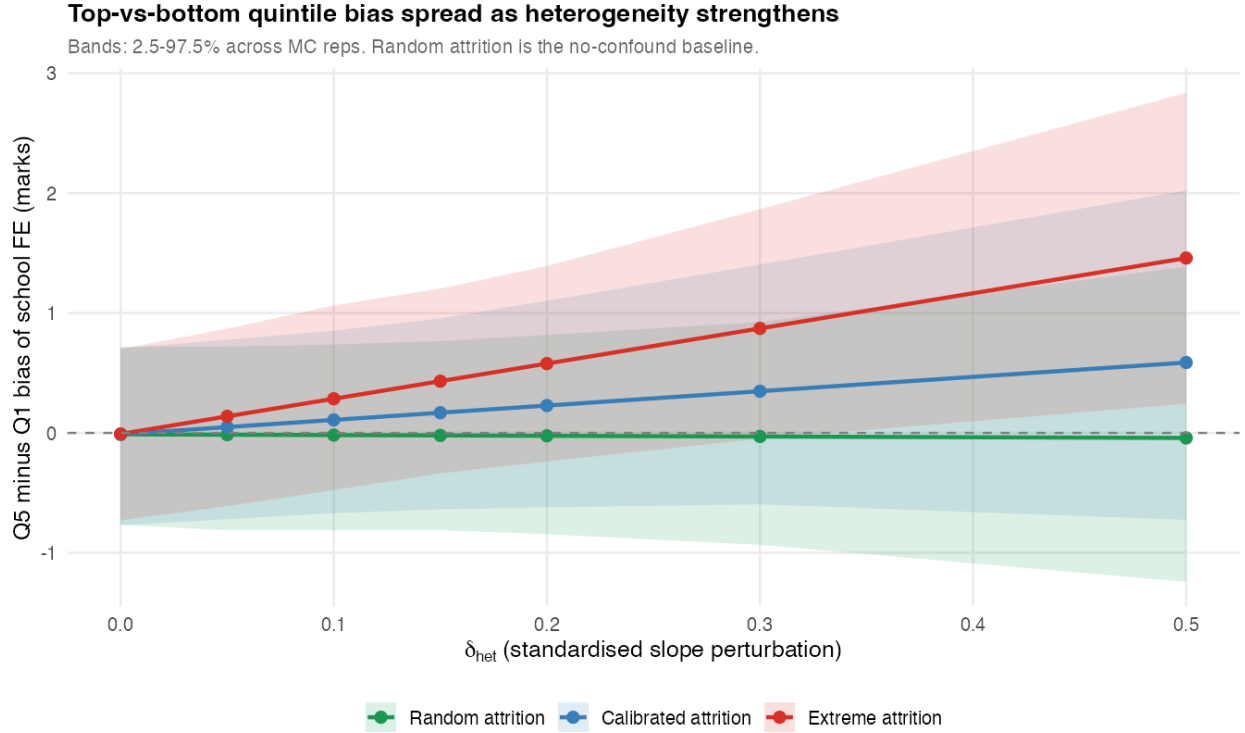


Figure 2: Top-vs-bottom quintile bias spread of the TWFE school FE as a function of the heterogeneity parameter δ_{het} , under three of the Monte Carlo attrition scenarios (Table 6 of the main text). Random attrition is the no-confound baseline; calibrated attrition matches the observed $\text{cor}(\text{linkage rate, school quality}) \approx 0.021$; extreme attrition jointly imposes strong selection on ability and school quality. Bands are 2.5–97.5 percentile across 300 simulations per cell. *Sample*: simulated data calibrated to the Gauteng core, 2018 and 2022 cohorts.

The implication for the school-FE estimand is that the calibrated-attrition $Q_5 - Q_1$ bias spread crosses 1 mark only at $\delta_{\text{het}} \approx 0.85$ (linear extrapolation from the simulated grid), corresponding to a +1 SD ability student at a +1 SD school receiving roughly 10.1 extra marks of learning beyond the linear-additive prediction. Even under the adversarial extreme scenario the threshold is not crossed until $\delta_{\text{het}} \approx 0.34$. For interaction strengths $\delta_{\text{het}} \in [0, 0.30]$, a +1 SD/+1 SD premium of up to 3.6 marks, the calibrated $Q_5 - Q_1$ bias spread stays at or below 0.35 marks, and the rank correlation with the truth at or above 0.955.

The δ_{het} values at which the bound is crossed are large by the standards of the within-school dispersion the data exhibit, and would require a substantively bigger interaction than the heterogeneous-pass-through diagnostic (Appendix H.3) or the population-invariance diagnostic in Table 22 detect. The threat is therefore bounded: under the empirically calibrated attrition severity, heterogeneous treatment effects do not move the school-FE quintile bias spread by more than half a mark for any plausible interaction strength.

Empirical-anchor bound. Four additional cells extend the scenario list to span selection severity at $1\times$, $2\times$, $5\times$, and $10\times$ the empirically observed correlation between school linkage rates and estimated school effects ($r = 0.021$, Table 5). These cells set $\delta_1 = 0$ (no selection on individual ability) and calibrate δ_2 to hit each target correlation, isolating the school-correlated selection channel that the empirical IPW diagnostic measures. Across the resulting 4×7 sub-grid (four K cells $\times$ seven δ_{het} values), the maximum absolute $Q_5 - Q_1$ bias spread is 0.118 marks, achieved at the most adversarial cell ($K = 10\times$, $\delta_{\text{het}} = 0.5$), well under 2% of $\sigma_\tau \approx 6$ marks. At the empirically anchored cell ($K \approx 1\times$, realised $r = 0.02$) the bound tightens to 0.052 marks across the full δ_{het} sweep. The interpretation is that the school-FE differential is essentially immune to selection acting only on school quality, even at adversarial multiples of the empirical correlation: the heterogeneous-TE channel only bites materially when selection also truncates the within-school ability distribution (the $\delta_1 > 0$ scenarios in the main grid above), and the empirical IPW shift of -0.005 marks rules that out at empirically realistic severities.

Caveats. The bound is conditional on the DGP. The interaction is parameterised as a *single* slope perturbation in standardised tau; richer forms of heterogeneity (non-monotone, threshold, or driven by an unobservable other than tau itself) are not in the simulation’s support. The replication code documents the full parameterisation.

Connected-set coverage by province

Identification requires that the bipartite network of Grade 7 schools and Grade 8 classes be connected; the largest connected component within each province-year cell defines the estimating frame (§2.4 of the main text). Table 7 reports coverage by province for the 2018 cohort (the 2022 cohort shows near-identical results). Coverage is high across all six provinces, confirming that virtually all schools and students enter the estimating frame. The Herfindahl–Hirschman index (HHI) measures concentration in Grade 8 intake patterns; values below 0.6 indicate moderate-to-low concentration. See §3.3 of the main text for the network-structure discussion.

Table 7: Connected-set coverage, 2018 cohort. The 2022 cohort shows near-identical results.

Province	Gr 7 Schools	% Schools Connected	% Students Connected	Mean HHI
EC	3117	98.8	99.3	0.525
GT	1432	99.4	99.8	0.273
LP	2151	99.6	99.9	0.479
MP	970	99.4	99.8	0.525
NW	834	99.9	100.0	0.402
ZN	2543	99.7	99.8	0.429

Sample: Province panel (6 provinces, estimated separately), 2018 cohort. Includes all Grade 7 schools and students matched to a Grade 8 destination. Mean HHI is the average Herfindahl–Hirschman concentration index across Grade 8 schools within each province: for each Grade 8 destination school, the HHI is the sum of squared shares of its intake from each Grade 7

feeder school. Values below 0.6 indicate moderate-to-low concentration. Table 7 reports the 2018 cohort; the 2022 cohort is near-identical.

Appendix B: Measurement Error Analysis

Attenuation bias, full specification table

The table below reports the Grade 7 mark coefficient (β_1) across eight specifications: an OLS baseline, two multi-term averages, two heaping robustness checks, two functional-form checks, and the structural de-heaping estimate. The baseline Term 4 coefficient on raw marks, 0.743, indicates substantial but incomplete persistence from Grade 7 marks to Grade 8 performance. Multi-term averages (rows 2–3) confirm that Term 4 is not appreciably noisier than the composite. Structural de-heaping of both mark distributions, reported in full in Table 10, shifts the Gauteng-2022 headline coefficient by only $\Delta\beta_1 = -0.0054$, confirming that removing the heaping mass does not materially move the persistence parameter.

Rows 4–5 address heaping directly: including heaping-point dummies or excluding observations at 30/40/50 moves the coefficient by at most 0.001. The piecewise-linear specification (row 6) reports the coefficient for the first segment only, and the quadratic specification (row 7) the linear term; both are functional-form diagnostics taken up in Table 9. Row 8 carries the Rubin-pooled Gauteng-2022 estimate after de-heaping both mark distributions; the full 12-cell de-heaping matrix is in Table 10.

Table 8: Attenuation bias: Grade 7 mark coefficients across specifications.

Specification	β_1	SE	N	Schools	Adj. R ²	Within R ²
(1) Baseline: Term 4 only	0.743	0.010	225,040	1,595	0.574	0.360
(2) Multi-term average (T1-T4)	0.774	0.006	225,040	1,595	0.579	0.369
(3) Multi-term average (T1-T3, excl. T4)	0.745	0.007	225,040	1,595	0.569	0.353
(4) Baseline + heaping-point dummies	0.742	0.011	225,040	1,595	0.574	0.360
(5) Baseline, excluding marks at 30/40/50	0.742	0.011	208,969	1,592	0.576	0.359
(6) Piecewise linear (breaks at 30, 40, 50)	-0.051	0.160	225,040	1,595	0.580	0.369
(7) Quadratic in Gr7 mark	0.072	0.064	225,040	1,595	0.581	0.370
(8) Structural de-heaping (Gauteng-2022, Rubin-pooled)	0.790	—	225,040	1,595	—	—

Sample: Gauteng core, pooled 2018 and 2022 cohorts; row 8 is Gauteng 2022 alone.

Piecewise vs linear, functional-form sensitivity

Table 8 row 6 reports a near-zero “leading coefficient” for the piecewise specification with breaks at the heaping thresholds. This subsection demonstrates that the result is an artefact of the basis parameterisation, not evidence of nonlinearity in the underlying Grade 7 to Grade 8 relationship.

Six specifications are estimated on the national six-province Mathematics panel: the baseline linear, the linear specification re-estimated on threshold-excluded sub-samples (dropping marks within ± 2 pp and ± 5 pp of 30 / 40 / 50), the current piecewise specification (knots at 30 / 40 / 50), a richer piecewise specification with knots every 5 pp across the heaping range, and a natural cubic spline whose internal knots are placed *away* from 30 / 40 / 50 by design. For each row we report the leading basis coefficient (the coefficient the regression reports) and the **effective slope**, defined as the population-average partial derivative of the fitted regression function with respect to the baseline mark. For the linear rows the two are identical; for the

multi-basis rows the effective slope is the comparable scalar (a mass-weighted average of segment slopes), while the leading coefficient is the slope within a single segment conditional on the other basis functions.

Two patterns close the question. First, the linear slope estimated on threshold-excluded sub-samples (0.69 at ± 2 pp; 0.705 at ± 5 pp) is *larger* than the full-sample baseline (0.685), not smaller. If the baseline slope were identified from heaping-induced mass movement, removing the heaping bands should collapse it towards the piecewise leading coefficient; instead it grows. Second, every multi-basis specification (piecewise at 10 pp bins, piecewise at 5 pp bins, and the natural cubic spline with knots placed away from heaping) returns an effective slope of 0.681, indistinguishable from the linear baseline. The reduced leading coefficients in Table 8 row 6 and in row (4) of Table 9 reflect three mechanical absorption channels in the multi-basis design: bin-specific intercepts soak up *all* between-bin level variation; the heaping spikes thin the within-bin support that the remaining slope is estimated on; and where threshold-adjacent students cluster at particular feeder schools, the bin dummies compete with the Grade 7 school fixed effect α_j for the same variation.

Table 9: Sensitivity of the Grade 7 mark coefficient to functional form.

Specification	β_1 (leading)	SE	Effective slope	N	Within R ²
(1) Baseline linear (full sample)	0.685	0.001	0.685	1,135,367	0.307
(2) Linear, excluding ± 2 pp of 30/40/50	0.690	0.001	0.690	793,887	0.314
(3) Linear, excluding ± 5 pp of 30/40/50 (keep < 26 or > 54)	0.705	0.002	0.705	528,466	0.274
(4) Piecewise linear, knots at 30/40/50	0.283	0.011	0.682	1,135,367	0.313
(5) Piecewise linear, knots every 5pp on [25,55]	0.084	0.019	0.681	1,135,367	0.313
(6) Natural cubic spline, internal knots at 15, 22, 60, 70 (away from 30/40/50)	—	—	0.682	1,135,367	0.313

Sample: National (6 provinces), pooled 2018 and 2022 cohorts.

Structural de-heaping: $\Delta\beta_1$ across 12 national cells

The table below reports the Rubin-pooled shift in the Grade 7 persistence parameter β_1 after structural de-heaping of both the Grade 7 baseline and Grade 8 endline mark distributions, separately for each of the 12 (province $\times$ cohort) cells. The maximum absolute shift across all cells is $|\Delta\beta_1| = 0.0154$ (2.1% of the worst-cell baseline β_1 , occurring in Eastern Cape 2018); the direction is negative in 12 of 12 cells. The de-heaping procedure fits a degree-7 polynomial counterfactual density excluding a ± 2 -mark window around 30, 40, and 50, separately for each mark and for each (province $\times$ cohort $\times$ quintile) cell, and draws $m = 25$ imputed mark panels with both marks redrawn where heaped. Rubin’s combining rules pool within-imputation analytical SEs with between-imputation variance.

Table 10: Structural de-heaping: $\Delta\beta_1$ per national cell (Rubin-pooled, $m = 25$ imputations).

Province	Cohort	$\hat{\beta}_1$ (raw)	$\Delta\beta_1$	$ \Delta /\beta_1$ (%)	$ \Delta $ / SE
Eastern Cape	2018	0.728	-0.0154	2.12	3.699
Gauteng	2018	0.761	-0.0079	1.04	2.423
Limpopo	2018	0.661	-0.0107	1.62	2.891
Mpumalanga	2018	0.701	-0.0089	1.27	2.093
North West	2018	0.694	-0.0076	1.10	1.492
KwaZulu-Natal	2018	0.675	-0.0085	1.26	2.537
Eastern Cape	2022	0.744	-0.0083	1.12	2.323
Gauteng	2022	0.795	-0.0054	0.68	1.879
Limpopo	2022	0.716	-0.0071	0.99	2.325

(continued)

Province	Cohort	$\hat{\beta}_1$ (raw)	$\Delta\beta_1$	$ \Delta /\beta_1$ (%)	$ \Delta $ / SE
Mpumalanga	2022	0.732	-0.0060	0.82	1.553
North West	2022	0.753	-0.0062	0.82	1.400
KwaZulu-Natal	2022	0.774	-0.0072	0.93	2.520

Sample: National 6-province Mathematics panel, 2018 and 2022 cohorts. Both marks de-heaped jointly. Sign-uniform: $\Delta\beta_1 < 0$ in all 12 of 12 cells. Economic shift $|\Delta\beta_1|/\beta_1 \leq 2.1\%$ in every cell.

Structural de-heaping: $\hat{\alpha}_j$ Pearson ρ across 12 national cells

The table below reports the Pearson correlation between raw and de-heaped school fixed effects $\hat{\alpha}_j$ under both-mark de-heaping, along with the mean and SD of the mark-scale shift in $\hat{\alpha}_j$, for each of the 12 (province $\times$ cohort) cells. The minimum ρ across all cells is 0.997.

Table 11: Structural de-heaping: $\hat{\alpha}_j$ Pearson ρ , raw vs de-heaped school effects.

Province	Cohort	Schools	ρ (min)	ρ (median)	Mean shift (marks)	SD shift (marks)
Eastern Cape	2018	3,079	0.99730	0.99758	+0.716	0.565
Gauteng	2018	1,424	0.99952	0.99958	+0.424	0.283
Limpopo	2018	2,143	0.99828	0.99841	+0.927	0.471
Mpumalanga	2018	964	0.99820	0.99846	+0.618	0.463
North West	2018	833	0.99845	0.99878	+0.596	0.429
KwaZulu-Natal	2018	2,535	0.99866	0.99878	+0.684	0.462
Eastern Cape	2022	3,129	0.99848	0.99861	+0.502	0.445
Gauteng	2022	1,479	0.99956	0.99963	+0.267	0.254
Limpopo	2022	2,236	0.99887	0.99894	+0.528	0.409
Mpumalanga	2022	1,047	0.99919	0.99932	+0.518	0.342
North West	2022	937	0.99897	0.99912	+0.317	0.412
KwaZulu-Natal	2022	3,062	0.99902	0.99910	+0.232	0.400

Sample: National 6-province Mathematics panel, 2018 and 2022 cohorts. ρ is the Pearson correlation between raw and de-heaped $\hat{\alpha}_j$ within each province $\times$ cohort cell. The mean shift is the change in each fit's zero, which follows from the change in the slope; it does not move schools relative to one another.

B.5: Heaping heterogeneity across school quintiles

Ranked by the DBE poverty quintile, a school-funding classification rather than a measure of academic standing, the 40-mark threshold dominates heaping severity in every quintile. The figure below reports excess students above the local-polynomial counterfactual at each threshold, expressed as a percentage of each quintile total; the 30% and 50% thresholds contribute negligibly across all five poverty quintiles. Ranked instead by academic standing, heaping is not flat: schools in the bottom fifth by average Grade-7 mark heap about 5.6 times as often as those in the top fifth, since low-scoring schools carry more students just below the 40 line. That gradient does not distort the recovered ranking ([Appendix L.5](#)).

B.6: IV exclusion check

An earlier draft of §5.3 of the main text considered an instrumental-variables strategy that would use Term 1 marks as an instrument for the Grade 7 Term 4 mark. Three diagnostics establish that the instrument is contaminated by the same threshold-pushing mechanism the IV is meant to correct; the IV strategy was therefore dropped in favour of the structural de-heaping reported in §5.3 of the main text. The three diagnostic figures are reproduced here for completeness.

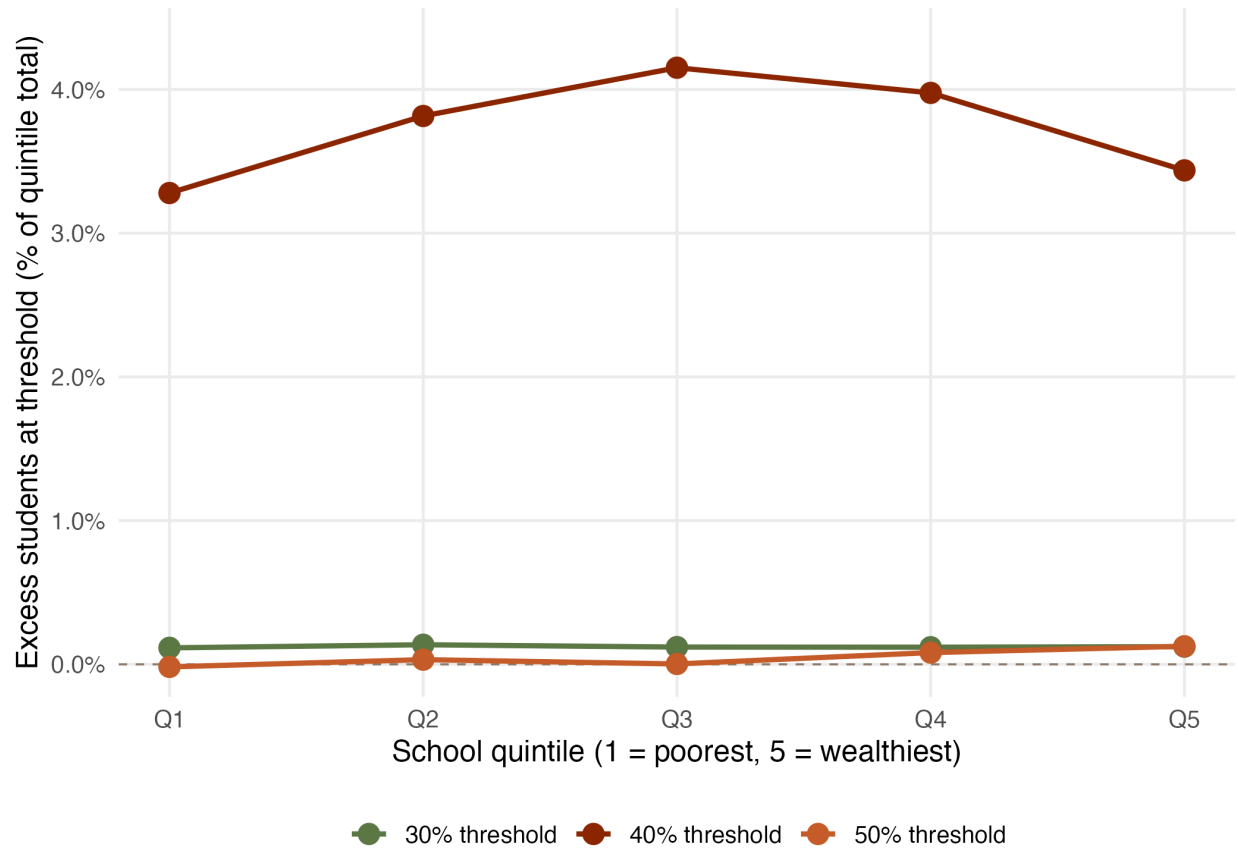


Figure 3: Excess students above the local-polynomial counterfactual at each threshold, expressed as a percentage of each DBE poverty quintile total (Q1 poorest to Q5 wealthiest, a funding classification). The 40% threshold dominates across all five poverty quintiles; the 30% and 50% thresholds contribute negligibly. Ranked by academic standing instead, heaping is heaviest at the lowest-scoring schools ([Appendix L.5](#)). *Sample:* National (6 provinces), pooled 2018 and 2022 cohorts.

Figure 4 documents that Term 1 marks themselves heap at 2.26 times the counterfactual frequency at mark 40. Figure 5 shows that the joint $(T_1, T_4) = (40, 40)$ cell occurs at 6.09 times its expected frequency under independence, so within-student heaping is correlated across terms. Figure 6 reports the school-level co-movement of T1 and T4 heaping intensities (OLS slope 0.30, Pearson $r = 0.27$), establishing that the heaping mechanism is a persistent within-school phenomenon rather than a term-specific idiosyncrasy.

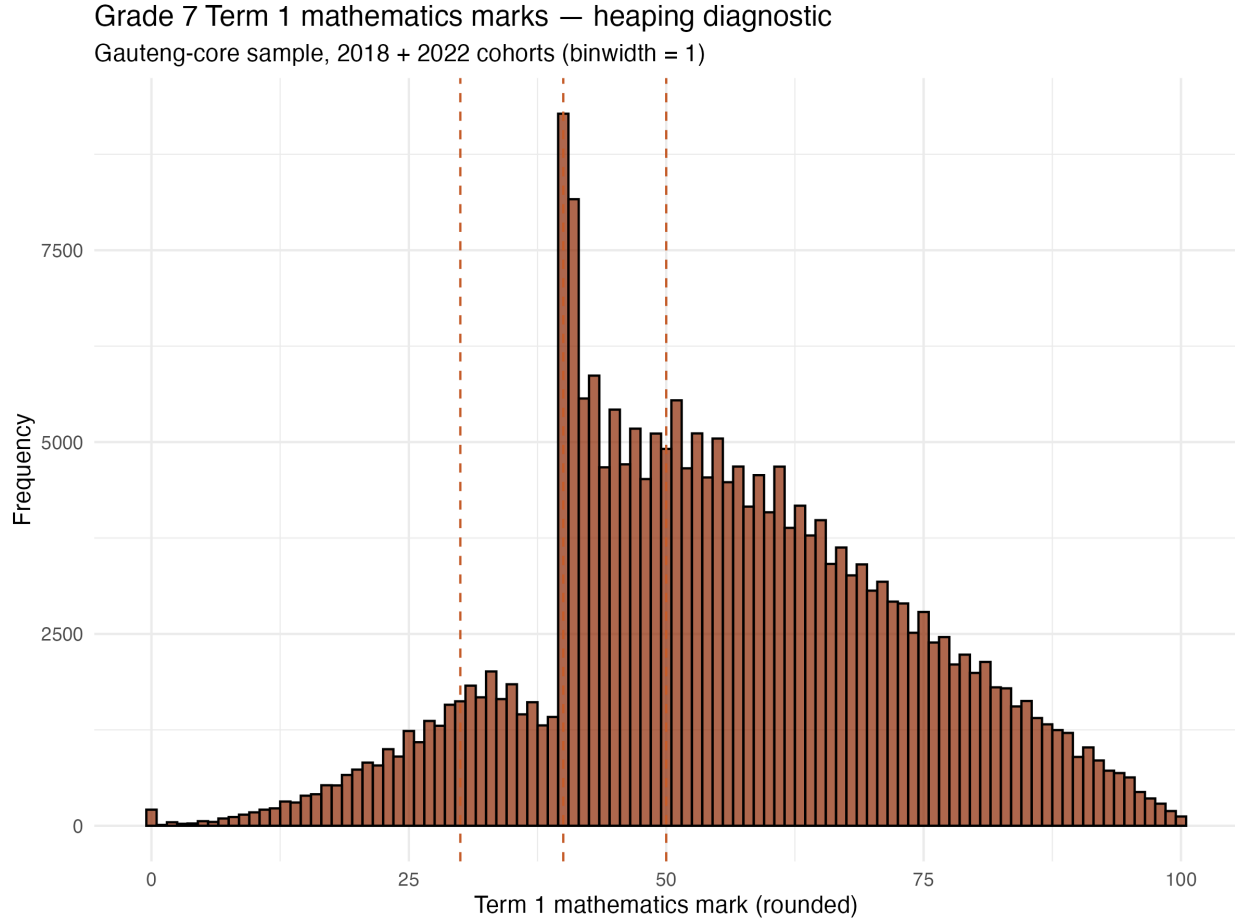


Figure 4: Term 1 mark distribution: excess mass at the 40% threshold (heap index 2.26). The IV exclusion restriction requires that Term 1 marks are uncontaminated by threshold-pushing; this figure documents that the restriction fails. *Sample:* Grade 7 mathematics pupil-term records, pooled 2018 and 2022 cohorts.

B.7: Monte Carlo recovery under threshold-pushing

The figure below reports the headline Monte Carlo exercise described in §5.1 of the main text. Panel A shows the baseline coefficient $\hat{\beta}_1$ under three estimators, namely oracle (computed on pre-heaping marks), OLS with heaped marks, and IV using Terms 1–3 as a DGP comparator, across heaping indices from 1 (no heaping) to 3 (beyond observed levels). Panel B reports the rank correlation between estimated and true school effects.

The generated marks in Figure 8 confirm that threshold-pushing is active in the DGP, set to a target heaping index of 2.4 at the 40-mark threshold. The spike at 40 accounts for 2.7% of heaped marks against 1.6% in the smooth reference. The bounding sweep in [Appendix L.2](#) carries realised severity well above the empirical Gr7-2022 index of 2.24, so it tests a live threat, not a phantom one.

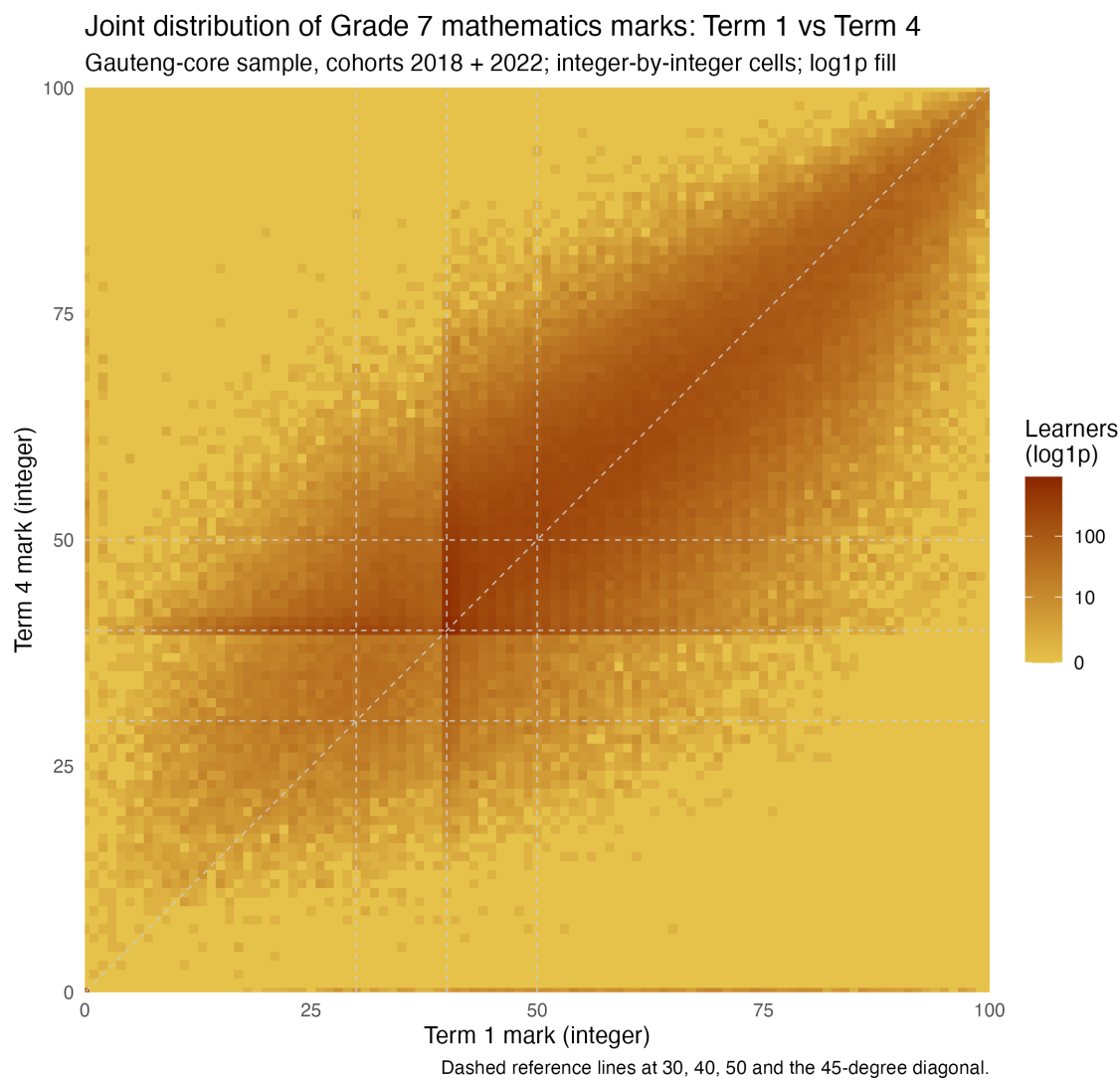


Figure 5: Joint distribution of Term 1 and Term 4 marks: the (40, 40) cell holds 6.09 times the mean count of its sixteen neighbouring cells, confirming correlated within-student heaping across terms. *Sample:* Grade 7 mathematics pupils observed in both Term 1 and Term 4, pooled 2018 and 2022 cohorts.

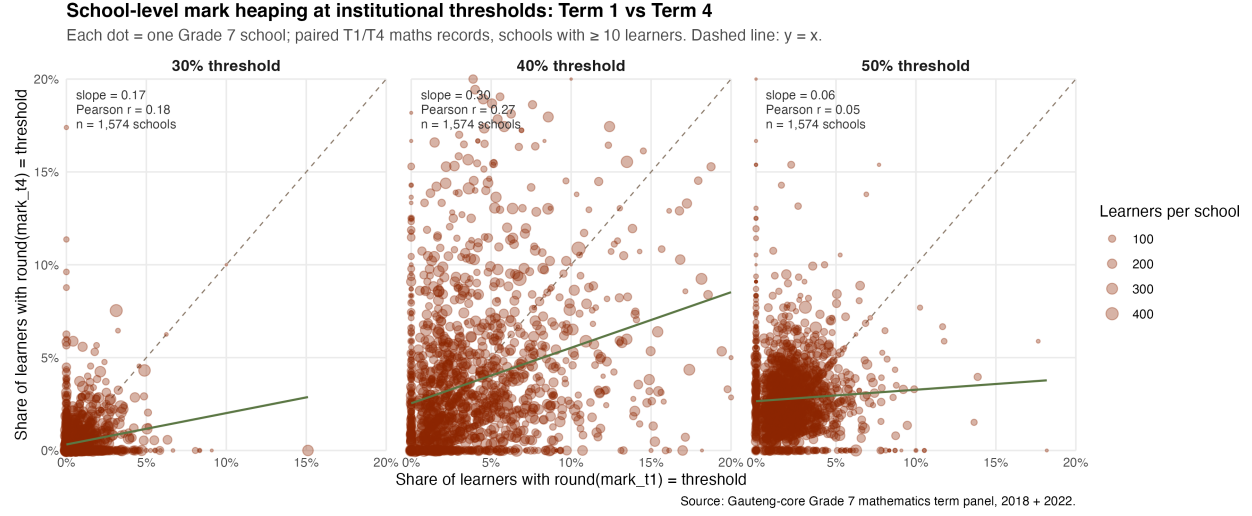


Figure 6: School-level Term 1 vs Term 4 heaping intensities: OLS slope 0.30, Pearson $r = 0.27$. Schools with stronger Term 4 heaping also show stronger Term 1 heaping, establishing that the heaping mechanism is a persistent school-level phenomenon, not a term-specific idiosyncrasy. *Sample:* Grade 7 mathematics schools with at least ten pupils, pooled 2018 and 2022 cohorts.

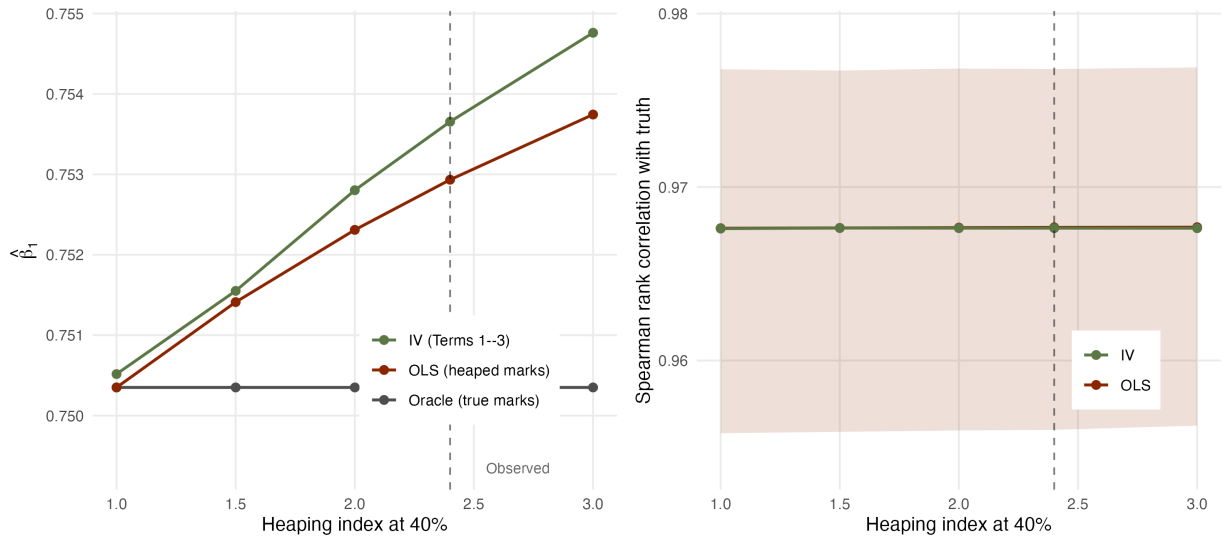


Figure 7: Monte Carlo: threshold-pushing in Grade 7 marks. Panel A: baseline coefficient under oracle, OLS, and IV. Panel B: Spearman rank correlation between estimated and true school quality. Dashed line marks the observed heaping index. 500 replications per scenario. *Sample:* simulated data calibrated to the Gauteng core, 2018 and 2022 cohorts.

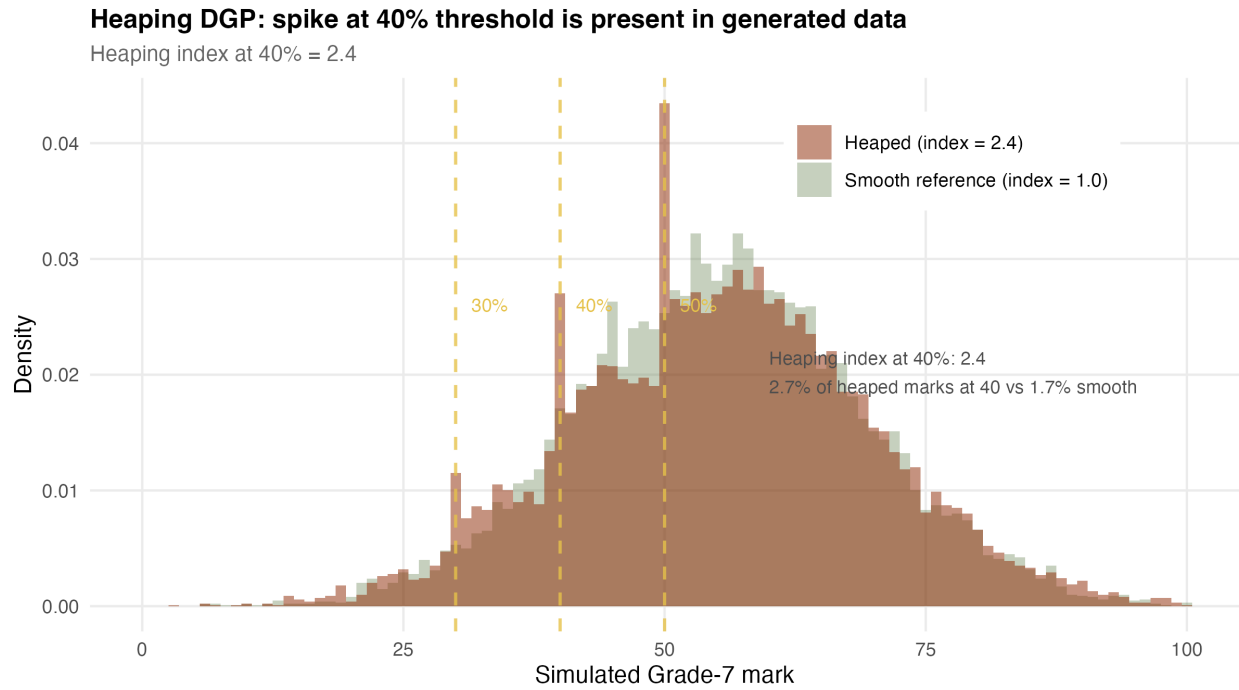


Figure 8: DGP face-validity: threshold-pushing in generated Grade 7 marks. Density of heaped marks (DGP setting: heaping index 2.4 at the 40-mark threshold) overlaid on the smooth unheaped truth, with vertical lines at the 30, 40, and 50 thresholds. The spike at 40 is visible: 2.7% of heaped marks land exactly at 40, against 1.6% in the smooth reference (a 1.7-fold excess). The empirical Gr7-2022 heaping index is 2.24; the bounding sweep in [Appendix L.2](#) carries severity well above that empirical level. *Sample*: generated data from the Monte Carlo design of §5.1 of the main text, not observed students.

Appendix C: Subject-Specific vs Pooled Estimation

The main analysis estimates school effects using Mathematics marks only. This appendix reports results from two alternative specifications – English First Additional Language (EFAL) only, and a pooled model that combines both subjects in a single TWFE regression – and explains why the main analysis uses Mathematics alone.

Why Mathematics only?

Grade 8 Mathematics and EFAL are taught in different classrooms, by different teachers, using different assessments. A school effect estimated from a single-subject model captures that subject’s instructional environment. A pooled model forces a single baseline coefficient (β_1) on two distinct production functions and produces a composite school effect that blends both subjects.

Three-way comparison of key statistics

Table 12: Key estimation statistics: Mathematics vs English vs Pooled.

Statistic	Mathematics	English	Pooled
N students (2018)	532,034	435,408	967,442
N students (2022)	664,853	543,522	1,208,375
Reliability (pooled across years)	0.985	0.974	0.974
Cross-year correlation (precision-weighted)	0.624	0.500	0.461
Cross-year correlation (disattenuated)	0.620	0.486	0.465
Split-sample SB reliability (2018)	0.939	0.810	n/a (national only)
Split-sample SB reliability (2022)	0.935	0.901	n/a (national only)
SES gradient (marks per quintile step)	1.93	0.73	1.92

Sample: Mathematics and English columns are on the Gauteng-core sample, pooled 2018 and 2022 cohorts. The Pooled column predates the Gauteng-core scoping decision: those values reflect a national stack of all six provinces, both subjects and both cohorts. The Pooled and subject columns are therefore not on a common sample and the three-way ordering is not directly informative; the split-sample row is reported as “n/a (national only)” because no comparable national split-sample variant of the subject pipelines exists. The reliability row pools the six provinces’ wedges, each measured from its own zero, so it is higher than the reliability within any one fit.

Sample-scope caveat: pooled vs Mathematics and English are not directly comparable

The Pooled column above reflects a national-sample stack of all six provinces, both subjects and both cohorts; the Mathematics and English columns are scoped to the Gauteng-core sample used as the headline sample (§2.4 of the main text). The pooled estimator was retired during that re-scoping, so no like-sample pooled run currently exists and the three-way disattenuated- r ordering in the table is not informative about a within-sample subject-aggregation effect.

Appendix D: Province-Specific Results

Province diagnostics and robustness

Per-province estimation diagnostics for both cohort years. Each row reports, for that (province $\times$ cohort) cell, the numbers of schools and students, the main-estimate β_1 , the median standard error of the estimated school effects, a reliability proxy, the signal-to-noise ratio and a summary verdict. The final two rows aggregate the per-cell estimates into an n -schools-weighted mean and an inverse-variance-weighted mean of β_1 . The pooled β_1 in these rows is the appropriate common-slope reference for the paper, in preference to reference-choice-contaminated national TWFE values. The lowest reliability in the six LCC-connected

provinces is 0.93; Gauteng achieves the highest signal-to-noise ratio, consistent with its larger class sizes and denser student-mixing networks; Eastern Cape has the lowest precision.

Table 13: Per-province robustness diagnostics.

Province	Cohort	Schools	Students	β_1	Median SE	Reliability	SNR	Verdict
EC	2018	3,079	108,606	0.500	0.122	0.936	3.82	Caution
EC	2022	3,129	117,770	0.526	0.123	0.929	3.61	Caution
GT	2018	1,424	106,514	0.601	0.067	0.980	7.00	High
GT	2022	1,479	135,852	0.592	0.067	0.980	6.99	High
LP	2018	2,143	89,601	0.533	0.100	0.953	4.48	Medium-High
LP	2022	2,236	116,910	0.580	0.095	0.957	4.74	Medium-High
MP	2018	964	63,557	0.585	0.082	0.966	5.33	Medium-High
MP	2022	1,047	75,922	0.607	0.083	0.969	5.57	Medium-High
NW	2018	833	43,038	0.593	0.096	0.961	4.93	Medium-High
NW	2022	937	56,439	0.610	0.091	0.966	5.34	High
ZN	2018	2,535	120,718	0.530	0.105	0.943	4.08	Medium-High
ZN	2022	3,062	161,960	0.585	0.103	0.948	4.27	Medium-High
Pooled, school-weighted	—	22,868	1,196,887	0.558	0.100	0.951	4.68	—
Pooled, inverse-variance	—	22,868	1,196,887	0.569	—	—	—	—

Sample: Province panel (6 provinces, estimated separately), 2018 and 2022 cohorts. See §2.4 of the main text for the Gauteng-core headline.

§4.4 of the main text compares the production function on raw marks and on the calibrated level across the six provinces, with province-by-cohort effects.

National replication of the calibration

Run on a national extract of the same school-based-assessment records, with the paper’s specification unchanged, the two-way calibration estimates the difficulty wedge, which the exhibits’ notes call the intake-school effect, in 31 Mathematics cells, each a province by Grade 7 cohort by Grade 8 anchor term (Table 14, Panel A). The cells span 3 cohorts, those of 2021 to 2023, and 8 provinces. The Grade 8 anchor in these cells is the Term 1 or the Term 4 mark; the paper’s own anchor is the Term 2 mark, which the national extract does not carry. Gauteng is not among those provinces: the national extract holds no Grade 7 Term 4 mark for Gauteng in these cohorts, so the province reaches no estimated cell. Panel B reproduces the paper’s own Gauteng-core estimates at the Term 2 anchor; they are not re-estimated here. The paper’s own extract carries more students per school than the national extract in the provinces where the extracts overlap, so sample sizes are not comparable between the panels and the comparison rests on slope and dispersion. Across the cells of Panel A the slope of the standardised Grade 8 anchor mark on the standardised Grade 7 mark runs from 0.475 to 0.770, and the dispersion, the cross-school standard deviation of the wedge, runs from 0.439 to 1.271 within-province standard deviations of the standardised mark. The lowest slope and the highest dispersion both come from a single thin cell, North West’s 2022 cohort at the Term 1 anchor, which rests on few feeder schools. The Gauteng core’s slope and dispersion in Panel B lie inside both ranges.

Table 14: **The estimator returns the same slope and dispersion in 31 out-of-sample cells.**

	Slope (1)	SD (2)	Med. SE (3)	SNR (4)	Sig. (%) (5)	R2 (6)	Learners (7)	Schools (8)
<i>Panel A. This extract, national, Mathematics</i>								
EC 2021 T1	0.678 (0.042)	0.627	0.091	6.895	66.7	0.503	13,239	6
EC 2021 T4	0.721 (0.044)	0.700	0.078	8.925	100.0	0.567	11,760	6
EC 2022 T4	0.705 (0.006)	0.835	0.096	8.694	78.2	0.513	22,142	435
EC 2023 T1	0.644 (0.006)	0.847	0.105	8.077	79.0	0.417	21,276	428
EC 2023 T4	0.737 (0.006)	0.896	0.092	9.720	77.8	0.552	21,695	451
FS 2023 T1	0.718 (0.013)	0.679	0.090	7.536	79.5	0.472	5,782	78
FS 2023 T4	0.770 (0.012)	0.744	0.084	8.823	88.0	0.571	5,565	83
KZ 2021 T1	0.550 (0.002)	0.450	0.101	4.458	63.1	0.333	179,591	3,394
KZ 2021 T4	0.670 (0.002)	0.505	0.093	5.430	68.5	0.466	180,005	3,401
KZ 2022 T1	0.594 (0.002)	0.460	0.101	4.577	64.7	0.372	184,627	3,459
LP 2021 T1	0.529 (0.003)	0.439	0.100	4.398	63.4	0.324	112,221	2,228
LP 2021 T4	0.630 (0.002)	0.470	0.086	5.487	69.7	0.473	113,393	2,243
LP 2022 T1	0.554 (0.003)	0.455	0.098	4.659	65.0	0.336	113,952	2,246
LP 2022 T4	0.669 (0.003)	0.517	0.087	5.949	72.3	0.493	76,877	1,583
LP 2023 T1	0.544 (0.008)	0.522	0.084	6.176	72.5	0.374	10,648	207
LP 2023 T4	0.705 (0.007)	0.591	0.071	8.311	82.3	0.585	10,750	204
MP 2021 T1	0.637 (0.007)	0.681	0.087	7.870	79.0	0.411	16,835	252
MP 2021 T4	0.720 (0.006)	0.682	0.075	9.071	77.5	0.534	16,815	249
MP 2022 T1	0.596 (0.003)	0.502	0.086	5.828	69.2	0.364	69,632	1,017
NC 2021 T1	0.546 (0.006)	0.598	0.083	7.230	76.0	0.397	18,371	300
NC 2021 T4	0.654 (0.005)	0.605	0.076	7.957	77.9	0.531	17,950	299
NC 2022 T1	0.609 (0.006)	0.572	0.085	6.735	72.9	0.415	19,480	306
NW 2021 T1 †	0.579 (0.027)	0.720	0.108	6.660	87.5	0.396	2,918	24
NW 2021 T4 †	0.579 (0.027)	0.721	0.108	6.664	87.5	0.396	2,920	24

(continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
NW 2022 T1 †	0.475 (0.019)	1.271	0.100	12.747	76.5	0.274	2,878	51
NW 2022 T4 †	0.588 (0.020)	1.118	0.095	11.815	77.5	0.401	2,714	49
NW 2023 T1	0.713 (0.013)	0.668	0.081	8.202	86.0	0.593	3,091	50
NW 2023 T4	0.761 (0.012)	0.654	0.072	9.091	76.5	0.669	3,126	51
WC 2022 T1	0.482 (0.040)	0.589	0.189	3.118	55.1	0.202	1,008	49
WC 2022 T4	0.575 (0.039)	0.701	0.177	3.960	54.2	0.300	1,003	48
WC 2023 T4	0.736 (0.003)	0.507	0.064	7.944	77.6	0.593	54,417	890
<i>Panel B. The paper, Gauteng core, Grade 8 Term 2 anchor</i>								
GT 2018 T2	0.601 (0.003)	0.471	0.067	7.000	70.4	0.404	106,514	1,424
GT 2022 T2	0.592 (0.002)	0.468	0.067	6.990	74.8	0.398	135,852	1,479

Note. Column (1), Slope, is the slope of the Grade 8 anchor mark on the Grade 7 baseline mark, with its standard error in parentheses on the line beneath. Columns (2) to (5) describe the estimated intake-school effect in that cell: SD is its standard deviation across schools, Med. SE the median standard error of a single school's effect, SNR the signal-to-noise ratio those two imply, and Sig. (%) the share of schools whose effect is individually distinguishable from zero. Column (6), R2, is the within R-squared of the fitted cell; columns (7) and (8) are its learners and its Grade 7 feeder schools. Each row of Panel A is one province by cohort by anchor cell, estimated with the paper's own specification unchanged: a valid mark at both ends, a minimum origin-school size of ten, marks standardised within province by cohort by subject, the largest connected component, and cells below 100 rows skipped. The cell code gives the province, the Grade 7 cohort year and the Grade 8 anchor term. Effects are in within-province standard deviations of the standardised mark. Panel A holds the 31 cells that estimate, out of the 54 Mathematics cells the extract defines: 20 are structurally absent because one of the two mark blocks is not in the extract at all, and 3 are too thin to fit. A dagger marks a cell whose two anchors rest on the same underlying block, so its anchor comparison is not informative; 4 cells carry it. Levels are not comparable between the two panels: the paper's Gauteng-core sample carries roughly 5% more learners per school than this extract does, so only the shape and the slope of the comparison carry information, and no statement about one panel's level against the other's is licensed here. The standard error in column (1) is on the calibration slope and is clustered at the Grade 7 origin school; the standard error in column (3) is on a single school's effect and is the pipeline's own sigma-over-root-n convention. No cell of this table carries a significance star: column (5) reports the share of schools reaching significance rather than starring individual rows. Where stars appear elsewhere in this document they follow one ladder, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. A blank cell is structurally empty: every column but (1) is a cell-level quantity that belongs to the coefficient line rather than to the standard-error line beneath it. The missing-value glyph for this document is an em dash. None arises here. Source: analyses/sba_nsc_link/outputs/tables/tab3_main_estimates.csv, via outputs/tables/mathematics/tab_nsc_replication.csv, which records the source table and selector behind every cell. **Sample:** Panel A, Grade 7 Term 4 to Grade 8 Term 1 or Term 4 transitions in the national school-based-assessment extract, Mathematics, cohorts 2021 to 2023; the 31 estimated cells carry 1,003 to 184,627 learners and 6 to 3,459 feeder schools each. Panel B, the paper's own Gauteng-core Mathematics cells at the Grade 8 Term 2 anchor, 2018 and 2022 cohorts, 1,424 to 1,479 feeder schools.

The shape of the national dispersion

Figure 9 plots the distribution of the wedge across feeder schools for the national extract’s Grade 7 2021 cohort at the Term 1 anchor, beside the distribution for the Gauteng core’s Grade 7 2018 cohort at its Term 2 anchor. The national extract’s distribution covers 6,204 feeder schools in 6 provinces, and the Gauteng core’s covers 1,424 schools. The spreads are nearly the same. The cross-school standard deviation of the wedge is 0.467 in the national extract and 0.471 in the Gauteng core, each about half a within-province standard deviation of the standardised mark. The national extract’s distribution pools the provinces with an estimated cell for that cohort at that anchor; the remaining cells of Table 14 are not drawn. Both distributions are centred within province, so the comparison is of their shape. Sample sizes are not comparable between the distributions, because the paper’s own extract carries more students per school than the national extract where the extracts overlap.

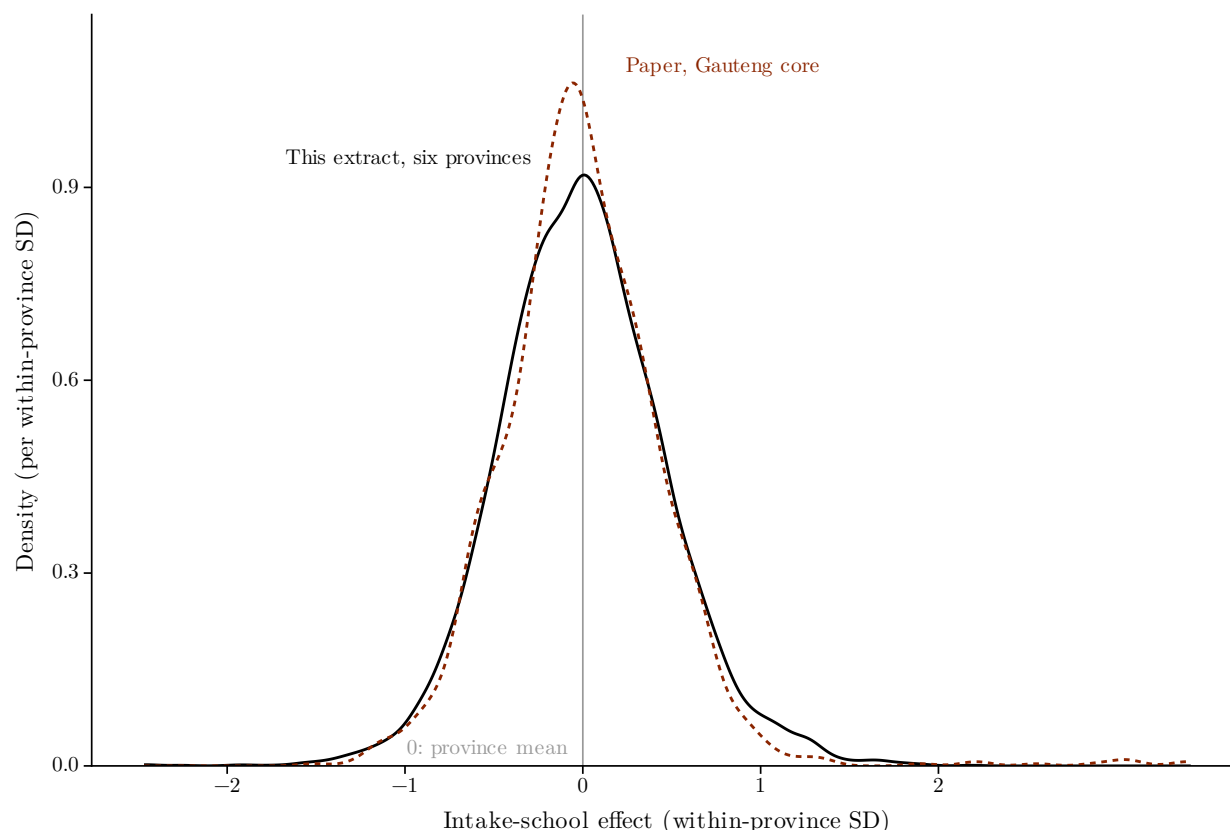


Figure 9: **The national extract and the paper’s cohort show the same spread.**

Note. The two curves are kernel densities of the estimated intake-school effect, in within-province standard deviations of the standardised mark. The black solid curve is this extract’s Grade 7 2021 cohort at the Grade 8 Term 1 anchor, Mathematics, 6,204 feeder schools in 6 provinces, standard deviation 0.467. The crimson dashed curve is the paper’s own Grade 7 2018 Gauteng-core cohort at its Grade 8 Term 2 anchor, 1,424 schools, standard deviation 0.471. The grey vertical line at zero is the province mean, drawn behind the data: both series are centred within province, which is what makes them comparable in shape. **These are two cells, not a picture of the 31 cells of the replication table above.** Levels are not comparable between them: the paper’s Gauteng-core sample carries roughly 5% more students per school than this extract does, so only the shape of the comparison carries information. Source: `analyses/sba_nsc_link/outputs/estimates/alpha_hat_2021_t1_math.parquet` and `outputs/tables/school_learning_estimates.csv`, via `outputs/tables/mathematics/tab_nsc_alpha_dispersion.csv`. **Sample:** the 6,204 Mathematics feeder schools of the Grade 7 2021 cohort estimated at the Grade 8 Term 1 anchor across 6 provinces, against the 1,424 Gauteng-core feeder schools of the paper’s Grade 7 2018 cohort.

The Grade 8 anchor term

Table 15 reports how closely the wedge estimated at a Grade 8 Term 1 anchor agrees with the wedge estimated at a Grade 8 Term 4 anchor, for feeder schools of the national extract estimated at both anchors in the same cohort and province. In the pool of cells from the Grade 7 2021 and 2023 cohorts, the ordering of schools is closely preserved: the precision-weighted Pearson correlation between the wedges from the Term 1 and Term 4 anchors is 0.902 over 6,676 feeder schools. Corrected for the sampling error in both sets of estimates, the correlation is 0.942. Appendix H.8 also moves the Grade 8 anchor, with a different question and a different sample. That subsection re-estimates the wedge at Term 2, Term 3 and Term 4 anchors on the Gauteng Mathematics panel and examines whether the no-fee gap in the wedge widens as the anchor moves later in the year. Table 15 moves the anchor between Term 1 and Term 4 on the national extract and shows whether the ordering of schools is preserved. Panel B of Table 15 reports the standard deviation of the wedge at each of the anchors of Appendix H.8, taken from that subsection’s estimates.

The Gauteng estimation sample allows a sharper version of the check, because it carries all four Grade 8 term marks for the same students. If a primary school’s effect is all difficulty, its wedge at Grade 8 Term t is its offset scaled by that term’s persistence, $\alpha_j(t) = \lambda(t) s_j$, while the slope estimates $\lambda(t)\rho$. The wedges at any two terms should then correlate perfectly once estimation noise is removed, and their spreads should stand in the ratio of the two slopes. A learning channel that fades or builds through Grade 8 breaks the second prediction; one confined to a single term breaks the first. I re-estimate the baseline specification with each Grade 8 term as the outcome and remove estimation noise with the published split halves, dividing the covariance of two terms’ wedges across halves, which share no student, by the covariance of each term’s wedge across halves. Over the six pairs of terms and both cohorts the corrected correlation is at least 0.976. In 2018 the spread tracks the slope, departing from it by at most 3% at any term relative to Term 2. In 2022 it narrows through the year faster than the slope does: relative to Term 2, the wedge is 7% wider than the slopes predict at Term 1 and 7% narrower at Term 3. Part of a primary school’s effect in 2022 therefore fades through Grade 8, in the same order as the wedge, and the ordering is unaffected. The extract’s cohorts reach Grade 8 in 2022 and 2024, when the Term 4 report mark is in effect a year mark that contains the Term 1 mark, so its two anchors share a component; in Gauteng the corrected Term 1 against Term 4 correlation is 0.989 in 2018, before that change, and 0.985 in 2022, so the shared component does not raise the agreement.

Table 15: **The Grade 8 anchor term does not change the ordering of schools.**

	Headline pool (1)	Sensitivity pool (2)
<i>Panel A. This extract, Term 1 against Term 4 at the Grade 8 anchor</i>		
Pearson correlation	0.885	0.888
Pearson correlation, precision-weighted	0.902	0.903
Spearman rank correlation	0.902	0.900
Disattenuated Pearson correlation	0.942	0.942
Ratio of SDs, Term 4 over Term 1	1.083	1.084
Schools compared	6,676	8,551
<i>Panel B. The paper’s own Grade 8 anchor ladder, Gauteng core</i>		
SD of the wedge, Grade 8 Term 2 anchor	0.452	
SD of the wedge, Grade 8 Term 3 anchor	0.433	
SD of the wedge, Grade 8 Term 4 anchor	0.491	
Ratio of SDs, Term 4 over Term 2	1.086	

Note. Column (1) is the pre-declared headline pool; column (2) the sensitivity pool, which adds every further cell where both anchors estimate. Each row of Panel A is one statistic computed on the pair of school effects a cell produces at the Grade 8 Term 1 anchor and at the Grade 8 Term 4 anchor. The precision-weighted Pearson correlation weights each school by the inverse of the summed sampling variances of its two estimates; the disattenuated Pearson correlation divides by the geometric mean of the two reliabilities. The last row of

Panel A is a count of schools and the only row of the table that is not a correlation or a ratio, which is why decimals are uniform within a row rather than down the column. The exercise was declared before it was run, against four thresholds: a precision-weighted Pearson correlation of at least 0.80, a Spearman correlation of at least 0.80, a disattenuated Pearson correlation of at least 0.90, and a ratio of standard deviations inside 0.80 to 1.25. The source records the outcome as ANCHOR-INVARIANT, with every one of the 5 declared tests passing in both pools. Panel B is a different test of a neighbouring question and the two must not be read as one test run twice. Panel A moves the anchor between Term 1 and Term 4 on the national extract and asks whether the ordering of schools survives; the paper’s own anchor-horizon check moves its Gauteng Mathematics anchor from Term 2 to Term 3 to Term 4 and asks whether the no-fee gap widens. Different sample, different object, same direction: the anchor term does not change the reading. Levels are not comparable between the two panels: the paper’s Gauteng-core sample carries roughly 5% more learners per school than this extract does, so only the shape and the slope of the comparison carry information. No cell of this table carries a significance star: every entry is a correlation, a ratio or a count, and the source reports no p-value for any of them. Where stars appear elsewhere in this document they follow one ladder, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. A blank cell is structurally empty: the sensitivity pool is defined on Panel A alone, and the paper’s ladder in Panel B has no second pool. The missing-value glyph for this document is an em dash. None arises here. Source: analyses/sba_nsc_link/outputs/tables/tab3_anchor_t1_vs_t4.csv, via outputs/tables/mathematics/tab_nsc_anchor_invariance.csv, which records the source table, the pool and the named column behind every cell. **Sample:** Panel A, Mathematics feeder schools estimated at both the Grade 8 Term 1 and the Grade 8 Term 4 anchor in the same cohort and province; the headline pool is 6,676 schools over 7 province by cohort cells of the Grade 7 2021 and 2023 cohorts, and the sensitivity pool 8,551 schools over every further estimable cell. Panel B, the paper’s own Gauteng-core Mathematics panel of 2,562 school-cohorts carrying all four anchor marks.

Appendix E: Replication

Code repository

A public replication repository containing the R scripts required to reproduce every table and figure in the paper and this Online Appendix, conditional on access to the underlying administrative data, is available from the author on request and will be deposited at a permanent archival DOI.

Data access

The analysis uses confidential administrative records from the South African School Administration and Management System (SA-SAMS), the system in which schools capture their students’ records, in an extract covering six provinces. These data contain individual student records including school enrolment, assessment marks, and demographic information. Access requires a formal data-sharing agreement with the relevant provincial Department of Education.

Researchers wishing to replicate the analysis should contact:

- **Gauteng Department of Education:** Research Directorate
- **Department of Basic Education:** Education Management Information System (EMIS) Directorate

The data-sharing agreement governing this research prohibits redistribution of individual-level records. The replication package provides all code and intermediate aggregated outputs that do not compromise student confidentiality.

Appendix F: Sorting Permutation Test Methods

This appendix describes the test for non-random sorting of students from Grade 7 feeder schools into Grade 8 destinations, reported in §3.3 of the main text: a regression-based falsification test in the spirit of Rothstein

(2010) paired with a permutation-based reference distribution following the approach of Tartova (2025), who builds an empirical null from permutations of actual teacher-to-classroom assignments preserving each teacher’s classroom count (one-sided teacher-workload margin), adapted here to preserve the destination-specific flow counts from each feeder (full method below). The test probes an assumption of the empirical strategy: that within-grade assessment conditions do not differ systematically for students from different feeder schools.

Specification

For each (feeder school $j \times$ cohort year t) cell with at least 10 students and at least two distinct Grade 8 destinations, we fit the one-way ANOVA

$$z_{ijt} = \mu_{jt} + \sum_{k \in K_{jt}} \delta_{jtk} \mathbb{1}[d_i = k] + \nu_{ijt}$$

where z_{ijt} is a predetermined student characteristic, d_i is the Grade 8 destination, and K_{jt} is the set of destinations that received at least one student from feeder j in cohort t . The F-statistic for the joint null $H_0 : \delta_{jtk} = 0$ produces a cell-level p-value p_{jt} . Stacking p-values across all eligible cells yields an empirical distribution on $[0, 1]$ that, under random within-feeder assignment, should be approximately uniform.

We test three predetermined characteristics:

1. **Grade 7 Term 4 mark:** the headline continuous measure of student ability. Grade 7 marks are not comparable *across* feeder schools (that is the motivation for the calibration procedure itself), but they are meaningfully comparable *within* a feeder, which is exactly the conditioning the test uses.
2. **Gender:** binary indicator.
3. **Over-age status:** binary indicator for students aged 14 or older at the start of Grade 8.

We would prefer to use Grade 6 marks as the headline predetermined characteristic (a one-grade-earlier lagged measure of student attainment, taken before any Grade 7 sorting could affect the test) but they are not available in our data extract. (Rothstein (2010) uses the prior-year *gain*, ΔA_{ig-1} , in standardised end-of-grade test scores rather than course marks; “marks” is our setting-specific framing of the same idea, since standardised end-of-grade test scores are not available for South African Grade 6 cohorts.) School poverty quintile cannot be tested because it is a property of the feeder school itself and is therefore invariant within each cell.

Placebo construction

A common concern with cell-level F-tests on bipartite mover networks is that in cells with few students or few destinations, even under genuinely random assignment, the F-test’s finite-sample behaviour may depart from the asymptotic null, because the uniform p-value distribution assumed by the test does not hold in sparse cells. We adapt the placebo-calibration approach of Tartova (2025) to our setting. Where Tartova addresses finite-sample concerns regarding between-year classroom-composition overlap by building a null from permutations of actual teacher-to-classroom assignments (preserving each teacher’s classroom count), our adaptation addresses sparse-cell F-test behaviour by permuting student destinations within each feeder $\times$ cohort cell while preserving the destination-specific flow counts n_{jtk} . This uses a different margin (destination-side rather than teacher-side) but relies on the same core logic of building the null from the data’s own assignment structure.

Concretely, we permute the vector of Grade 8 destinations $(d_i)_{i=1}^{n_{jt}}$ *without replacement* within each feeder $\times$ cohort cell. This preserves the marginal feeder-to-destination flow counts n_{jtk} exactly (every destination still receives the same number of students from this feeder as in reality) but breaks the student-level matching. We repeat this permutation $R = 200$ times, recomputing the F-test for every characteristic $\times$ cell on each replicate. The resulting placebo distribution provides the reference against which the observed distribution is compared.

Summary statistics

For each predetermined characteristic, we report:

- The share of cells with $p < 0.05$ in the observed data and in the placebo distribution (the latter averaged across replicates, with 2.5/97.5 percentile bounds across replicates).
- The difference between the observed and placebo shares (the “excess”, in percentage points).
- The Kolmogorov–Smirnov distance of the observed p-value distribution from $U[0, 1]$, compared to the placebo mean KS distance.

Sorting permutation test, observed vs placebo

Table 16: Within-feeder permutation test for non-random sorting into Grade 8 destinations.

Characteristic	Cells	Share of cells with $p < 0.05$				KS distance	
		Observed	Placebo	95% interval	Excess	Observed	Placebo
Grade 7 Term 4 mark	22,142	0.2282	0.0596	[0.0568, 0.0621]	+16.9	0.239	0.018
Gender	22,039	0.0548	0.0136	[0.0121, 0.0149]	+4.1	0.220	0.169
Over-age	21,927	0.0976	0.0395	[0.0373, 0.0419]	+5.8	0.153	0.068

Excess is the observed share minus the placebo share, in percentage points, and the Grade 7 Term 4 mark is the headline characteristic. KS distance is the Kolmogorov–Smirnov distance of the p-value distribution from the uniform. Sample: National (6 provinces), pooled 2018 and 2022 cohorts, Mathematics. N cells tested: 22,142. Placebo distribution: 200 within-feeder permutations preserving feeder-to-destination flow counts.

Observed vs placebo p-value distributions

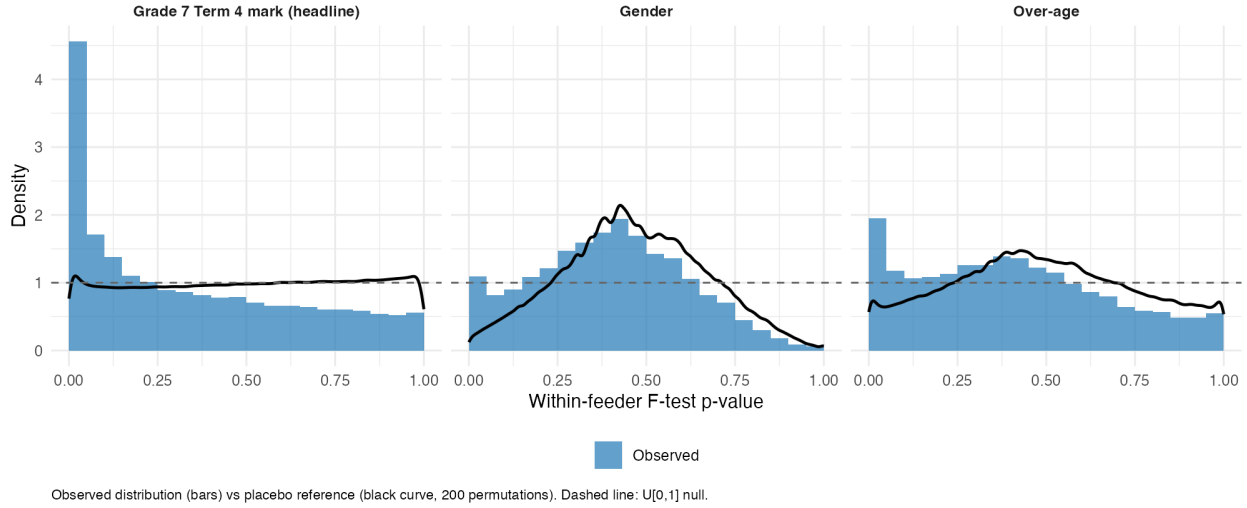


Figure 10: Within-feeder F-test p-value distributions for three predetermined characteristics, observed (bars) vs placebo reference (black curve, 200 permutations). Dashed line: uniform null. *Sample: as Table 16.*

Complementarity with the within-destination test

The test described here is the *within-feeder* complement of a *within-destination* test. The two tests probe the same underlying assumption from opposite sides of the bipartite Grade 7 $\rightarrow$ Grade 8 network:

- The within-feeder test (this appendix) asks: *within a feeder, do students with different predetermined characteristics sort into different destinations?*

- The within-destination test asks: *within a destination, do students from different feeders carry different main-equation residuals?*

Both tests are observable correlates of the same unobservable concern (within-destination differential treatment of feeder subgroups), and a coherent empirical story requires consistency between them.

Formal bound on sorting bias

The permutation test above probes the sorting threat empirically. This sub-section develops the *analytical* counterpart: a closed-form ceiling on how much non-random sorting into Grade 8 destinations could in principle move $\hat{\alpha}_j$, derived from the observed feeder-to-destination flow shares and the within-province cross-class variance of the destination fixed effects. The bound is referenced from §3.3 of the main text (the sorting threat).

Set-up

Let student i from Grade 7 feeder j be assigned to Grade 8 destination $k = k(i)$ by an unspecified rule, and write the destination fixed effect as $\gamma_k = \bar{\gamma} + u_k$ with $\mathbb{E}[u_k] = 0$ and $\text{Var}(u_k) = \sigma_\gamma^2$. The residual in the main text’s estimating equation decomposes as

$$\varepsilon_i = (\theta_i - \mathbb{E}[\theta_i \mid j, \text{Gr7Term4}_i]) + (\gamma_{k(i)} - \bar{\gamma}_j) + \xi_i,$$

where θ_i is latent end-of-Grade-7 ability and $\bar{\gamma}_j = \mathbb{E}[\gamma_{k(i)} \mid j]$ is the average destination quality faced by feeder j ’s leavers. Writing π_{jk} for the share of feeder j ’s linked leavers placed in destination k , this average is

$$\bar{\gamma}_j = \sum_k \pi_{jk} \gamma_k.$$

Bias expression

Partiallying out Gr7Term4_i and the destination fixed effects from the projection of $\hat{\alpha}_j$ onto the data, the asymptotic bias of $\hat{\alpha}_j$ relative to a reference feeder j_0 is

$$\text{plim}(\hat{\alpha}_j - \alpha_j) = \bar{\gamma}_j - \bar{\gamma}_{j_0} + O(\text{network correction}).$$

The network correction is second-order in the bipartite-graph density and is dominated by the $\bar{\gamma}_j - \bar{\gamma}_{j_0}$ term in any sample where the connected-set coverage is close to one (connected-set coverage exceeds 98.8%). The bias is therefore the difference between j ’s and the reference school’s average destination-class quality, after partiallying out students’ Grade 7 marks and the destination fixed effects.

Conditions for the bias to vanish

The strongest sufficient condition is *orthogonality*:

$$\mathbb{E}[u_k \mid j, \text{Gr7Term4}_i] = 0 \quad \text{for all } j.$$

Under orthogonality, $\bar{\gamma}_j = \bar{\gamma}$ for every feeder and the bias is exactly zero.

A weaker *bounded-heterogeneity* condition replaces the strict zero with a uniform bound, $|\mathbb{E}[u_k \mid j]| \leq B$ for some constant B , which turns the bias into a bounded error.

The operative condition in this paper is *averaging*: as the number of destinations per feeder grows, $\text{Var}_j(\bar{\gamma}_j) \rightarrow 0$, regardless of whether orthogonality holds at the individual destination. This is the condition the Herfindahl bound below makes quantitative.

Herfindahl bound

Treating $\{u_k\}$ as drawn from a distribution with within-province cross-class variance σ_γ^2 , a Lindeberg-type argument applied to the weighted sum $\bar{\gamma}_j = \sum_k \pi_{jk} (\bar{\gamma} + u_k)$ yields

$$\text{Var}(\bar{\gamma}_j | j) \leq \sigma_\gamma^2 \cdot H_j, \quad H_j = \sum_k \pi_{jk}^2.$$

The square root, $\sqrt{\sigma_\gamma^2 \cdot H_j}$, is a worst-case standard-deviation ceiling on the sorting-bias term for feeder j . The bound has the right limits: a feeder that places all leavers in a single destination has $H_j = 1$ and faces the full σ_γ ; a feeder that spreads leavers uniformly across K destinations has $H_j = 1/K$ and faces a bound shrinking as $\sigma_\gamma/\sqrt{K}$.

The plug-in estimate replaces σ_γ^2 by the within-province cross-class variance of the estimated destination fixed effects from the production specification, and replaces π_{jk} by the observed feeder-to-destination flow shares in the analysis panel.

Numerical ceiling

Table 17 reports the per-(province $\times$ cohort) summary: the within-province cross-class SD $\hat{\sigma}_\gamma$, the cross-school SD $\text{SD}(\hat{\alpha}_j)$, the median per-feeder Herfindahl H_j , the resulting median bound on $|\hat{\alpha}_j - \alpha_j|$ in standardised marks, and the median bound expressed as a fraction of $\text{SD}(\hat{\alpha}_j)$. The headline cell (Gauteng 2022, the same cell used elsewhere in §3 of the main text) shows a median ceiling of 0.263 standardised marks, equivalent to 56% of the cross-school SD of estimated school effects.

Table 17: Analytical Herfindahl ceiling on the sorting bias in $\hat{\alpha}_j$, by province and cohort.

Province	Cohort	Feeders	Destinations	SD(γ)	SD($\hat{\alpha}$)	Median H	Bound	Bound / SD
EC	2018	3,079	1,721	0.700	0.467	0.583	0.535	1.14
EC	2022	3,129	1,422	0.670	0.446	0.521	0.483	1.08
GT	2018	1,424	901	0.578	0.471	0.256	0.292	0.62
GT	2022	1,479	839	0.518	0.468	0.259	0.263	0.56
LP	2018	2,143	1,421	0.653	0.450	0.501	0.462	1.03
LP	2022	2,236	1,381	0.579	0.449	0.510	0.414	0.92
MP	2018	964	664	0.647	0.439	0.606	0.504	1.15
MP	2022	1,047	645	0.612	0.465	0.517	0.440	0.95
NW	2018	833	455	0.550	0.475	0.522	0.398	0.84
NW	2022	937	471	0.511	0.486	0.513	0.366	0.75
ZN	2018	2,535	1,840	0.676	0.426	0.444	0.451	1.06
ZN	2022	3,062	1,879	0.610	0.440	0.446	0.408	0.93

$\hat{\sigma}_\gamma$ is the within-province cross-class SD of destination fixed effects from the production specification. $\text{SD}(\hat{\alpha}_j)$ is the cross-school SD of estimated school effects. Median bound is $\text{med}_j \sqrt{\hat{\sigma}_\gamma^2 \cdot H_j}$ in standardised marks, expressed both absolutely (Bound) and as a fraction of $\text{SD}(\hat{\alpha}_j)$ (Bound / SD). The share of feeders whose bound sits below 0.1 standard deviations of $\hat{\alpha}_j$ is zero in every province-by-cohort cell, so that column is not shown. Sample: National (6 provinces), 2018 and 2022 cohorts, Mathematics. Bound construction: see [Appendix F](#).

Per-feeder distribution

Relationship to the empirical test

The bound and the permutation test are *independent* lines of evidence on the same unobservable. The Herfindahl bound is an analytical worst-case ceiling: it is what the bias could be if every u_k were chosen adversarially against every feeder’s destination portfolio. The permutation test is a non-adversarial empirical

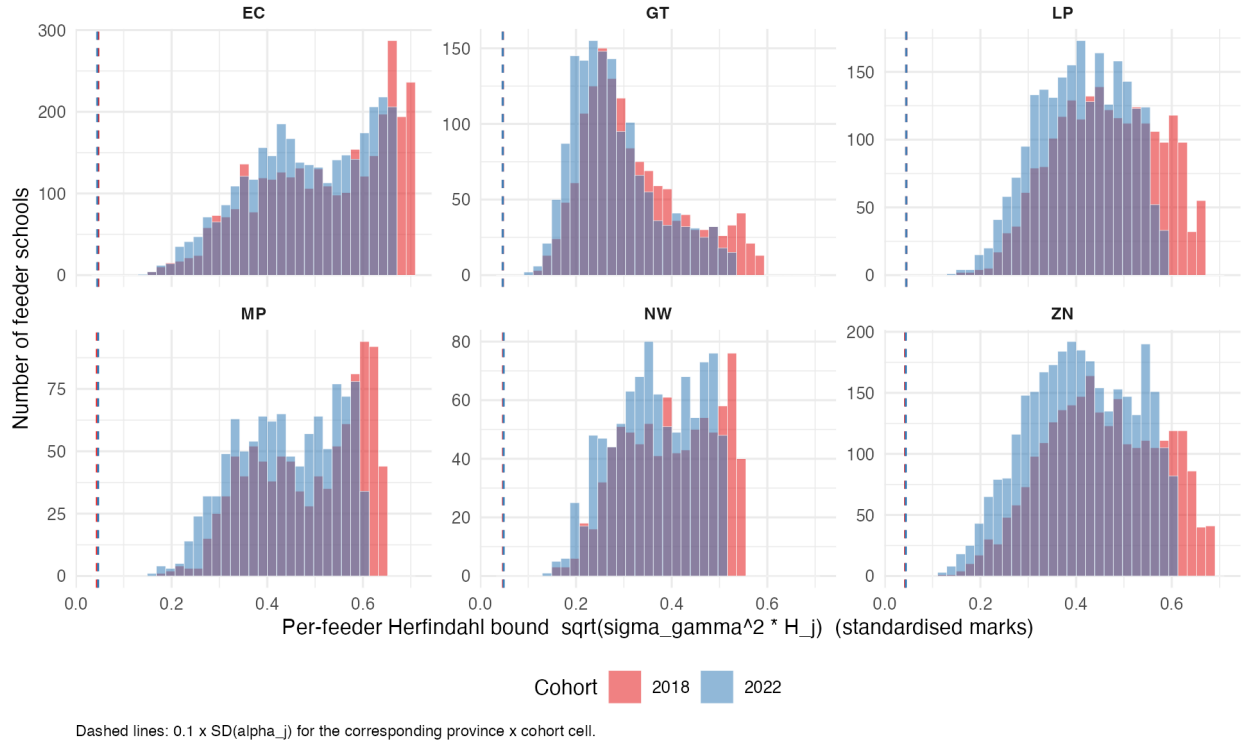


Figure 11: Per-feeder analytical sorting-bias ceiling $\sqrt{\hat{\sigma}_\gamma^2 \cdot H_j}$, in standardised marks, by province. Dashed lines mark $0.1 \times \text{SD}(\hat{\alpha}_j)$ for the corresponding province $\times$ cohort cell. *Sample:* National (6 provinces), 2018 and 2022 cohorts, Mathematics.

probe: it asks whether students with different predetermined characteristics actually sort across destinations within a feeder. A reader who wants reassurance under both the worst-case (H_j -based) and the typical-case (permutation) interpretations can read both off this appendix; the two are complements rather than substitutes for the sorting threat of §3.3 of the main text.

Appendix G: Cross-Cohort Stability of the Wedge

§4.2 of the main text reports the precision-weighted cross-year correlation as a falsification of the pure-noise null and otherwise points to this appendix. We collect the full set of cross-cohort diagnostics here. The estimand of the paper is purely cross-sectional: each year is calibrated against its own Grade 8 anchor, and there is no inter-temporal anchor that links calibrated learning levels across years. The cross-year correlations reported below are therefore *incidental empirical regularities*, not properties guaranteed by construction. They are reported to rule out the pure-noise null; they are not, and cannot be in this design, evidence of inter-temporal score comparability. The sub-unity gap between the observed correlations and 1.0 conflates four candidate sources (genuine cohort change, drift in assessment difficulty, estimation noise, and cross-cohort network composition); the term marks allow a partial decomposition, given at the end of this subsection. We bound one channel (network composition) below to show that it does not dominate; the remaining three are jointly unresolvable in this design, and the paper’s contribution does not depend on resolving them.

The precision-weighted Pearson cross-year correlation between the 2018 and 2022 wedges is 0.624 (Table 18), spanning a four-year gap that includes the COVID-19 disruption. We report Pearson and Spearman versions side by side: the Pearson statistic provides continuity with the school-level comparator in Liu et al. (2025) (and, at a greater interpretive distance, the teacher-level Chetty et al. (2014) benchmark); the Spearman rank correlation speaks to the paper’s substantive interest in school *rankings*.

Table 18: Cross-year stability.

Metric	Value
Schools in both cohorts	1,280
Raw correlation	0.602
Precision-weighted correlation	0.624
Disattenuated correlation	0.620
Rank correlation (Spearman)	0.601
% schools changing quintile	61.3%

Sample: Gauteng core, 2018 and 2022 cohorts.

The raw correlation of 0.602 rises only marginally to 0.624 after precision-weighting and to 0.620 after disattenuation for estimation error, confirming that the modest cross-year association is not driven by imprecisely estimated schools dragging the correlation towards zero. The Spearman rank correlation (0.601) tells a similar story in ordinal terms.

To benchmark this figure, we compare with the within-year split-sample reliability estimates reported in Table 26 and discussed in §4.2 of the main text. The Spearman-Brown reliability, the expected correlation between two independent estimates based on the full sample within the same year, is 0.935 in 2022 and 0.939 in 2018. The ratio of the cross-year correlation to the within-year reliability is 0.644 (2022) and 0.642 (2018), implying that about 36% of the within-year stable variance is not preserved across the four-year gap. This is a meaningful limitation: the measure is most informative as a within-cohort instrument. The term marks narrow the candidates without settling them. Re-estimating the wedge with each Grade 7 term as the baseline and each Grade 8 term as the anchor, and removing estimation noise with the split halves, the wedge’s cross-cohort correlation is 0.65: about 66% of its variance recurs in the other cohort. The cohort-specific remainder is shared by all four Grade 8 terms, so it is not the anchor term’s assessment. In 2018 most of it is specific to the primary school’s Term 4 mark, which wedges calibrated from the same school’s

Term 1 to Term 3 marks do not share; that part is about 22% of the wedge’s variance. (In 2022 the Term 4 mark is built from the earlier terms’ marks, so the comparison is not informative.) How hard a school’s year-end assessment is in a given year and what its students learn late in Grade 7 both fit; the design cannot tell them apart. The destination effect is far less a lasting property of the school: about 34% of its noise-free variance recurs in the other cohort, and about 41% is specific to a single Grade 8 term in a single cohort.

A middle cohort

The 2020 Grade 7 cohort sits between the paper’s two. Calibrated under the same specification against its 2021 Grade 8 anchor, it lets the same statistic be computed on the same schools over two two-year gaps as well as the four-year one. As throughout this appendix, the statistic is the correlation of the school estimates $\hat{\alpha}_j$ across cohorts. The four-year rung reproduces Table 18: in Gauteng the Spearman correlation between 2018 and 2022 is 0.600 over 1,210 schools, against 0.601 there. The two-year rungs are 0.620 from 2018 to 2020 and 0.746 from 2020 to 2022, over 1,136 to 1,203 schools.

The correlation does not decay with the length of the gap. Two years after 2018 it is already 0.620, within a few hundredths of its four-year value, and the four-year correlation sits well above 0.462, the product of the two two-year correlations, which is what steady year-on-year drift in each school’s assessment difficulty would produce. The stable variance is lost between cohorts rather than accumulated over elapsed time. That pattern fits a school component that persists across all three cohorts plus a component specific to each cohort. Pooled across six provinces the ordering is the same at a lower level: 0.467, 0.560 and 0.430 over 7,250 to 7,847 schools. The 2020 estimates carry two limits. They come from a separate run of the calibration, with each school required to have at least 20 students in both cohorts of a pair, and the 2020 cohort’s Grade 7 and Grade 8 years both fell in the COVID-19 disruption to schooling, which may account for part of the gap between the two two-year rungs.

Literature comparison

Having ruled out the pure-noise null above, the cross-year correlation admits an order-of-magnitude comparison with analogous correlations reported for school- and teacher-level measures in the literature. We offer such a comparison as a sanity check on our number’s order of magnitude, not as a validation: each cited estimand differs from ours in ways spelled out below, and a co-ordinated number is merely reassuring rather than corroborative. As a Pearson correlation, the figure is invariant to the linear rescaling that separates the two cohorts’ independently normalised scales and so can be benchmarked mechanically. The closest comparator is Liu et al. (2025) (school-level effects in the United States), who report GPA value-added cross-year correlations of 0.27 to 0.86 across settings. A more distant comparator is Chetty et al. (2014) (individual teacher effects rather than schools), whose cross-year autocorrelations are 0.30–0.48 at one-year lags and approximately 0.30 at four-year lags (Table 2 panel A, elementary mathematics). A correlation of 0.624 between our 2018 and 2022 cohorts (separated by four years and spanning the COVID-19 disruption) sits within the school-level range and slightly above the four-year teacher-level benchmark.

We report this number for context only. The cross-year gap conflates at least four distinct sources: genuine change in cohort composition, possible shifts in assessment difficulty, sampling noise, and changes in the bipartite feeder-to-destination network across cohorts that mechanically alter the anchor against which each school is measured. The second of these channels, drift in the underlying assessment instrument over time, is anticipated by the formal equating literature: Kolen and Brennan (2014, sec. 9.8 Maintaining Score Scales) notes that “the content of a test might slowly evolve, with the cumulative effect being that forms used in one year might have somewhat different content than forms used a few years later,” so that even well-designed equating relationships drift across years. Drift of that kind accumulates with elapsed time, and the middle cohort shows little accumulation: the two-year correlation from 2018 to 2020 is nearly as low as the four-year one (Appendix G), so gradual drift in the instrument is unlikely to be the main source of the gap. Our cross-sectional estimand does not require us to disentangle these sources, and we make no claim about which dominates. We bound the network-composition contribution in the next subsection: conditioning on the upper tertile of cross-cohort destination overlap, the correlation rises to 0.724. The within-cohort estimates remain interpretable on their own terms regardless of the cross-year figure.

Network-composition diagnostic

Because each Grade 7 school’s wedge is identified against the Grade 8 destinations its students reach, cross-year movement in $\hat{\alpha}_j$ reflects in part the fact that the bipartite feeder-to-destination network itself changes between cohorts. We quantify this channel directly. Across the 1,280 Gauteng Mathematics schools with $\hat{\alpha}_j$ estimated in both 2018 and 2022, the median Jaccard overlap of destination sets is 0.286 (p25 = 0.222, p75 = 0.368; Table 19): the typical school shares only about 30% of its Grade 8 destinations across the two cohorts. Recomputing the precision-weighted cross-year correlation on this overlap-restricted sample gives 0.645 for the full sample and 0.724 for the upper tertile of Jaccard ($J \geq 0.333$, $n = 484$; Table 20). The gap of about 0.08 between the two correlations bounds the contribution of cross-cohort network composition to the cross-year imperfection: it is non-trivial but leaves the bulk of the gap attributable to the other three sources listed above. Figure 12 plots per-school Jaccard against the cross-cohort change in $\hat{\alpha}_j$.

Table 19: Destination-set Jaccard overlap across the 2018 and 2022 Grade 7 cohorts (unweighted and flow-weighted).

Percentile	Unweighted	Flow-weighted
p05	0.129	0.298
p10	0.163	0.376
p25	0.222	0.508
p50	0.286	0.626
p75	0.368	0.722
p90	0.454	0.809
p95	0.500	0.853
mean	0.301	0.606

Sample: Gauteng core, 2018 and 2022 cohorts.

Table 20: Cross-year correlation of school effects conditional on cross-cohort destination-set overlap.

Subset	Schools	Pearson r	Pearson r (pw)	Spearman r
Full sample	1,280	0.602	0.645	0.601
Top tertile of J ($J \geq 0.333$)	484	0.636	0.724	0.645
Top quartile of J ($J \geq 0.368$)	327	0.626	0.726	0.641

Sample: Gauteng core, 2018 and 2022 cohorts.

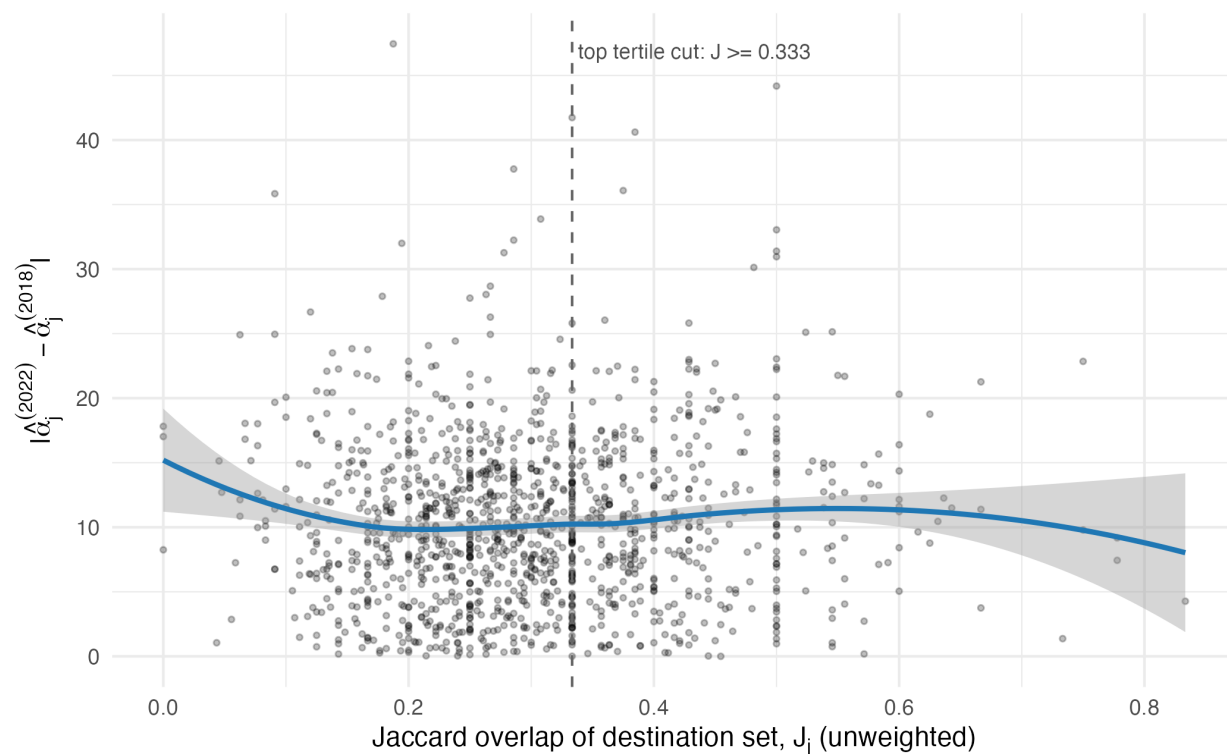
The observation that 61.3% of schools change quintile across cohorts is mathematically consistent with a correlation of 0.624 between the two years and does not represent additional evidence of instability beyond what the correlation already implies. Under a bivariate normal reference distribution with $r = 0.624$, the expected share of units remaining in the same quintile is approximately 37%, against an independence baseline of 20%. The observed share of 38.7% therefore sits ≈ 2 percentage points above the bivariate-normal prediction and ≈ 18.7 percentage points above the independence null, i.e. the quintile movement we observe is essentially the movement the reported correlation already implies, with no residual evidence of instability beyond it.

Why no contemporaneous cross-measure validation

The richer external-validation strategy that Liu et al. (2025) deploy (a horse race between their mark-based measure and a parallel test-based value-added measure built from state accountability assessments, with the near-zero correlation between the two interpreted as evidence that grade-based and test-based measures capture distinct underlying constructs) cannot be replicated in our setting. No contemporaneous

B4: cross-year movement vs. destination-set overlap

GT Mathematics, 1280 schools with $\alpha_{\hat{j}}$ in both cohorts



Source: B1_per_school.rds ($\alpha_{\hat{j}}$, clustered SE) + analysis_panel.parquet (destinations).

Figure 12: Per-school cross-cohort destination-set Jaccard against the cross-year change in $\hat{\alpha}_j$. Schools with thinner destination-set overlap show wider cross-year movement. *Sample:* Gauteng core, 2018 and 2022 cohorts.

census standardised test exists for the 2018 or 2022 Grade 7 cohorts: the Annual National Assessments were discontinued in 2014, and the survey-based instruments that remain (SACMEQ, TIMSS, PIRLS) are not student-linkable to the SA-SAMS records and cover different cohorts on different sampling frames. This is the principal external-validation gap in our setting and is a consequence of the developing-country data environment rather than a design choice; it is precisely why the within-year measurement-property exercises (§4.2 of the main text) and the Monte Carlo recovery (§5.1 of the main text) are made to carry the validation load that an external test would otherwise carry. The [Appendix I](#) comparison with Liu et al. (2025) returns to this point.

Appendix H: Additional Validation Diagnostics

This appendix collects auxiliary diagnostics referenced from §4.2 of the main text but not needed for the within-year reliability claims that section establishes.

H.1: Empirical Bayes shrinkage consistency check

Figure 13 plots the raw fixed-effects estimates against their empirical Bayes posterior means on the Gauteng Mathematics sample used in the reliability decomposition of Table 26. Under the parametric empirical Bayes estimator of Kane and Staiger (2008), the shrinkage weight applied to each school is $\lambda_s = \widehat{\text{Var}}(\alpha) / (\widehat{\text{Var}}(\alpha) + \text{SE}_s^2)$, where $\widehat{\text{Var}}(\alpha)$ is the signal variance of the school effects and SE_s^2 is the sampling variance of the individual estimate. Because the reliability ratios are 0.970–0.971, the shrinkage weight for the typical school is, by construction, approximately 0.97, and the empirical Bayes estimate must coincide with the raw fixed-effects estimate to within roughly 3% of the distance to the grand mean.

The figure is a *consistency check* on the pipeline rather than a comparison of two competing estimators: the fitted λ_s should track per-school precision, with smaller schools (higher SE_s) visibly pulled further towards the grand mean than larger schools. That pattern is what the figure displays. The equivalence is an arithmetic consequence of the within-cohort reliability reported in §4.2 of the main text, not an independent validation of it.

H.2: Observable placebo

The sensitivity-to-controls exercise in §4.2 of the main text asks whether $\hat{\alpha}_j$ moves when observable student characteristics are added to the specification. We complement it here with an observable placebo in the Chetty et al. (2014) style, transposed to the within-cohort setting. We regress Grade 8 Term 2 marks on a set of student-level SES proxies with no fixed effects to obtain a predicted score $\hat{A}_i^p$ for each student; we then aggregate $\hat{A}_i^p$ to the Grade 7 school level ($\bar{A}_j^p$) and regress $\bar{A}_j^p$ on $\hat{\alpha}_j$ across schools. Under the logic of Chetty et al. (2014), a near-zero slope would indicate that the school effect does not track sortable observable student characteristics. In our setting, where the estimand is a mark-scale offset that the design cannot separate from student selection, instructional effectiveness, and peer and contextual effects (§3 of the main text) rather than a causal teacher effect, the placebo is best read as a descriptive diagnostic of how much of the cross-school variation in $\hat{\alpha}_j$ co-moves with the SES-predictable component of Grade 8 performance.

Table 21: Observable placebo: regression of the SES-predicted school mean on the estimated school effect, by cohort and pooled (Gauteng Mathematics).

Cohort	Schools	Students	Slope	SE	CI lo	CI hi	Intercept	R sq.	SD(eff.) (marks)	SD($\bar{A}_p$) (marks)
2018	1,252	100,682	0.414	0.022	0.371	0.457	37.405	0.221	8.373	7.370
2022	1,312	129,886	0.425	0.014	0.397	0.453	41.308	0.402	8.522	5.712
Pooled	2,564	230,568	0.515	0.012	0.491	0.538	39.770	0.421	9.586	7.607

Sample: Gauteng core, 2018 and 2022 cohorts.

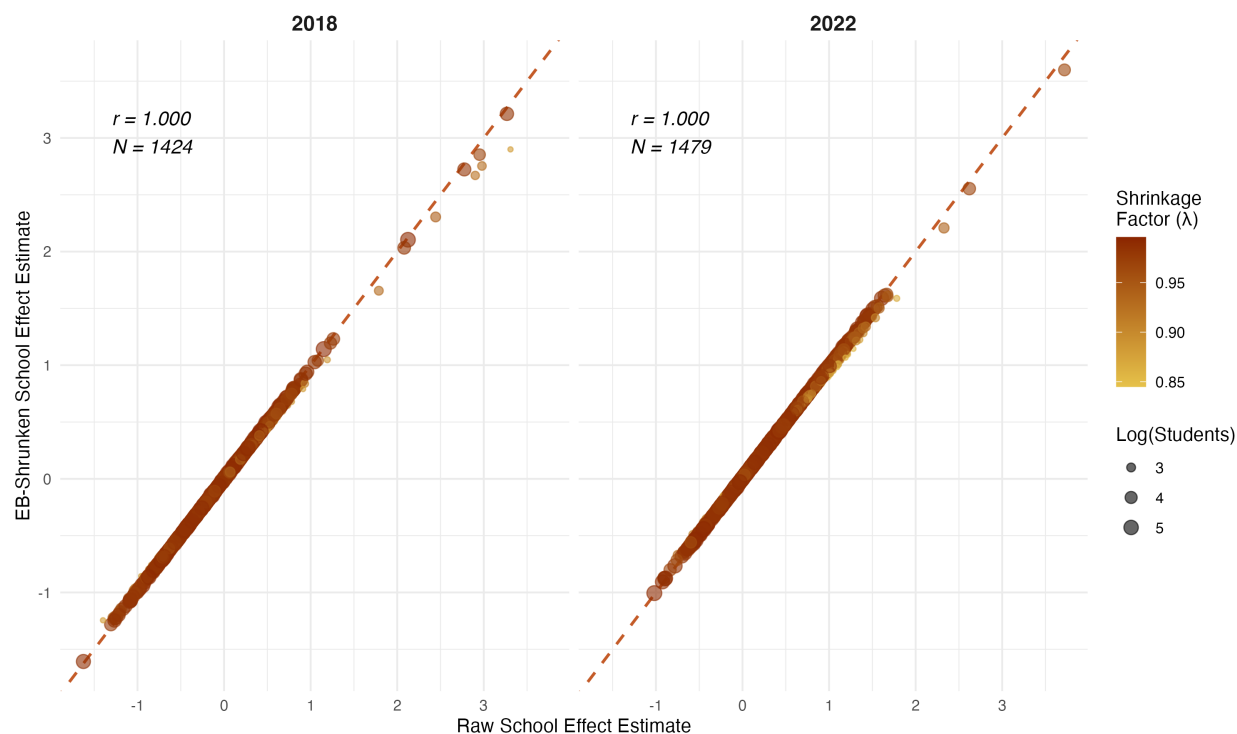


Figure 13: Raw fixed-effects estimates of the wedge, in SD of Grade 8 marks, versus empirical Bayes posterior means on the Gauteng Mathematics sample (same sample as the reliability decomposition in Table 26). Points cluster tightly around the 45-degree line because the within-cohort reliability of 0.97 forces the shrinkage weight λ_s to be close to 1 for the typical school. Point colour encodes λ_s and point size encodes $\log(n_students)$: the figure is a consistency check that λ_s tracks per-school precision in the expected way, not an independent comparison of two estimators. *Sample:* Gauteng core, 2018 and 2022 cohorts.

Figure B3: Observable placebo, within-cohort

Binned scatter (20 equal-count bins per cohort). A coefficient near zero implies the school effect does not pick

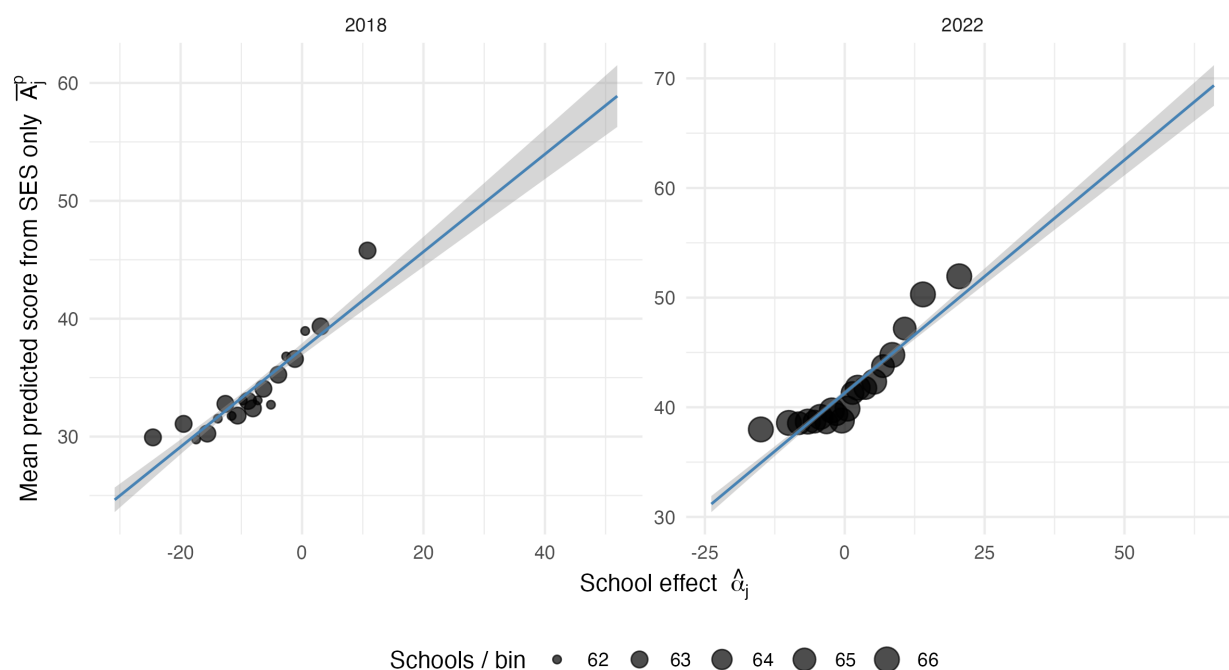


Figure 14: School-level scatter of the SES-predicted mean $\bar{A}_j^p$ against $\hat{\alpha}_j$, pooled across the 2018 and 2022 cohorts. *Sample:* Gauteng core, 2018 and 2022 cohorts.

Within each cohort, the slope of $\bar{A}_j^p$ on $\hat{\alpha}_j$ is 0.414 (2018) and 0.425 (2022). Pooling the two cohorts without a cohort term gives 0.515 (SE 0.012; 95% CI [0.491, 0.538]; $R^2 = 0.421$; $n = 2,564$ schools), higher because each cohort’s wedges are measured from their own zero. The slope is clearly positive: schools with higher $\hat{\alpha}_j$ are also schools whose students would be predicted to score higher from their SES / demographic profile alone. This is consistent with the socioeconomic gradient in §4.2 of the main text and with the fact that $\hat{\alpha}_j$ is an unconditional composite. The placebo does not constitute a failure of the measure under our cross-sectional estimand (it is exactly what a composite of student selection, instructional effectiveness, and peer and contextual effects built on a sortable population should look like), but it is a sharper statement of the sensitivity finding in §4.2 of the main text, and should be read alongside the quintile-conditioned specification reported there.

H.3: Population invariance across school subgroups

The cross-method and heterogeneous-pass-through checks in §4.2 of the main text test robustness of $\hat{\alpha}_j$ to functional form and heterogeneous slopes. A complementary diagnostic recommended by Kolen and Brennan (2014, Ch. 10) and formalised by Dorans and Holland (2000) is *population invariance*: the empirical content of a linking function’s invariance claim is that the function should not change materially when it is re-estimated inside non-overlapping subgroups of the population. We impose a single $\hat{\beta}_1$ on the entire Gauteng cohort; Table 22 reports the analogous check by re-fitting the baseline specification of §3.2 of the main text inside 12 cells defined by three school-level partitions (school mean Grade 7 mark quintile, national poverty quintile, and urban/rural status), alongside a school-size tercile check.

Table 22: Population-invariance diagnostic: $\hat{\beta}_1$ and subgroup-restricted $\hat{\alpha}_j$ re-estimated inside school subgroups.

Dimension	Level	N schools	β_1	SE	SD(signal) (marks)	Cor. w/ baseline
School mean Gr7 quintile	Q1	483	0.731	0.005	8.67	0.948
School mean Gr7 quintile	Q2	516	0.743	0.005	6.59	0.897
School mean Gr7 quintile	Q3	523	0.761	0.005	7.12	0.936
School mean Gr7 quintile	Q4	502	0.806	0.005	8.40	0.964
School mean Gr7 quintile	Q5	466	0.862	0.005	10.79	0.963
National poverty quintile	Q1	187	0.675	0.006	5.71	0.946
National poverty quintile	Q2	169	0.603	0.007	5.59	0.890
National poverty quintile	Q3	275	0.675	0.005	5.26	0.948
National poverty quintile	Q4	315	0.715	0.005	6.31	0.951
National poverty quintile	Q5	425	0.836	0.004	7.75	0.987
Urban / rural	RURAL	119	0.679	0.009	7.32	0.907
Urban / rural	URBAN	1,471	0.744	0.002	8.37	0.999
Full sample (baseline)	All schools	1,626	0.740	0.002	8.47	1.000

Each row re-fits the baseline specification on raw marks inside one subgroup, so $\hat{\beta}_1$ is Grade 8 mark points per Grade 7 mark point, not the within-province z-score slope of §3.2 of the main text. Sample: Gauteng core, 2018 and 2022 cohorts.

Across the 12 subgroup cells the within-stratum pass-through $\hat{\beta}_1$, in Grade 8 mark points per Grade 7 mark point, ranges from 0.603 to 0.862 (full-sample $\hat{\beta}_1 = 0.740$ on the same raw scale), a spread of 0.259. The spread is widest across the national poverty quintile (0.233) and the school mean Grade 7 quintile (0.132), with the top achievement and wealth strata showing the highest pass-through; it is narrowest across the urban/rural cut (0.065). Within each stratum, the subgroup-restricted $\hat{\alpha}_j$ tracks the baseline of §3.2 of the main text at Pearson $r \geq 0.890$ (Spearman $\rho \geq 0.874$) on the intersection of schools, with the looser correlations concentrated in the smallest cells where the Kane-Staiger noise share is mechanically inflated by the reduced within-school sample. Read against the school-size check ($r = 0.990$), the overall picture is one of partial invariance: the school ranking is reproduced across every cut but the *pass-through* is visibly

stronger at the top of the baseline-mark and wealth distributions, qualifying the single- $\hat{\beta}_1$ specification as a cohort-average summary rather than a *completely population invariant* linking function in Dorans and Holland (2000)’s sense.

H.4: Within-between variance decomposition

We sharpen the observation in §4.2 of the main text that the SES gradient leaves substantial within-quintile heterogeneity with a standard within/between variance decomposition of $\hat{\alpha}_j$ with respect to two coarse partitions: national poverty quintile (5 groups) and district (15 groups in Gauteng). 76% of the cross-school variance in $\hat{\alpha}_j$ is *within* poverty quintile, and 90% is *within* district (Table 23, Figure 15). This is the school-level analogue of the “most variation is within rather than between coarse groupings” result familiar from the teacher value-added literature (Rivkin et al. 2005), and it directly contradicts a naive sorting story in which the bulk of cross-school variation in $\hat{\alpha}_j$ is mechanically explained by SES or geography. The SES gradient in the marginal sense (a positive mean difference across quintiles) coexists with substantial dispersion *within* every quintile.

Table 23: Within- vs between-partition variance decomposition of school effects by cohort and partition (poverty quintile, district).

Cohort	Partition	N groups	Schools	Total var.	Between var.	Within var.	Share within
2018	quintile	5	1,252	70.114	11.139	58.975	0.841
2018	district	15	1,397	83.540	10.449	73.090	0.875
2022	quintile	5	1,312	72.631	23.640	48.991	0.675
2022	district	15	1,458	76.999	7.124	69.875	0.907
Pooled	quintile	5	2,564	71.386	16.860	54.526	0.764
Pooled	district	15	2,855	80.177	7.919	72.258	0.901

Sample: Gauteng core, 2018 and 2022 cohorts.

H.5: Subgroup reliability by baseline achievement

The within-year reliability exercises in §4.2 of the main text pool students across the achievement distribution and ask whether the recovered school effects are well-measured *on average*. This appendix asks the complementary question: is the measure equally precise across subgroups defined by baseline (Grade 7) achievement, and does the cross-school dispersion of recovered effects vary systematically with where students sit in the input distribution? The exercise mirrors the baseline-tertile heterogeneity analysis in Liu et al. (2025), and Appendix I returns to the contrast in framings.

For each cohort year we partition the linked Gauteng mathematics sample into quintiles of the Grade 7 baseline mark and re-estimate the main TWFE specification within each quintile, applying a per-slice school-size minimum of five students to suppress singleton-school cells. Table 24 reports the resulting signal-noise decomposition and the implied within-cohort reliability for each quintile, alongside the full-sample baseline.

Table 24: Subgroup reliability by Grade 7 baseline-mark quintile (Gauteng Mathematics).

Slice	Students	Schools	Components	Largest comp.	SD(signal) (marks)	Median SE (marks)	Reliability
2018 full	106,514	1,424	9	0.994	9.017	1.262	0.970
2018 Q1	20,451	874	10	0.990	6.587	2.030	0.861
2018 Q2	20,622	1,101	10	0.992	6.751	2.154	0.863
2018 Q3	20,714	1,196	14	0.989	7.528	2.367	0.873
2018 Q4	20,593	1,148	15	0.987	8.820	2.483	0.893
2018 Q5	20,381	963	17	0.984	10.523	2.513	0.917

(continued)

Slice	Students	Schools	Components	Largest comp.	SD(signal) (marks)	Median SE (marks)	Reliability
2022 full	135,852	1,479	6	0.997	8.746	1.241	0.971
2022 Q1	26,542	1,126	10	0.991	6.856	2.292	0.852
2022 Q2	26,654	1,267	9	0.994	8.249	2.383	0.892
2022 Q3	26,631	1,303	17	0.986	9.000	2.449	0.902
2022 Q4	26,682	1,283	11	0.992	9.868	2.536	0.913
2022 Q5	26,417	1,173	15	0.988	9.534	2.468	0.901

Sample: Gauteng core, 2018 and 2022 cohorts.

Two patterns emerge. First, within-quintile reliability is meaningfully lower than the pooled within-year reliability (0.85–0.92 across the ten cohort $\times$ quintile cells, against 0.970–0.971 in the full-sample baseline).¹ The mechanical reason is that each quintile slice contains roughly one-fifth of the students per school, which inflates the per-school standard error by a factor of approximately $\sqrt{5}$ and pushes the noise share of variance up. The substantive reading is that the measure remains informative for any subgroup but is most precisely measured when all students at a school are pooled, as the main pipeline does. Second, the dispersion of recovered school effects is largest for top-quintile students in both cohorts (SD of true effects rising from 6.6 marks at Q1 to 10.5 marks at Q5 in 2018, and from 6.9 to 9.5 marks in 2022). This is the *opposite* direction from the Liu et al. (2025) NC result on Grade 9 GPA, where the per-SD effect of high-VA middle schools is more than twice as large for bottom-tertile baseline-math students as for top-tertile (no such pattern in their MD sample), and [Appendix I](#) returns to the contrast.

H.6: Non-parametric estimation alternatives

The baseline and four alternative specifications of the wedge (quadratic, piecewise-linear, empirical-Bayes and cross-classified) give nearly the same wedges: each alternative correlates with the baseline at 0.987–0.998 (Pearson; §4.2 of the main text). Three further estimators, shown with them in Table 25, agree less because they estimate other objects.

Percentile rank equating (Method 5) gives each Grade 7 school the mean percentile rank of its students’ Grade 8 marks within their Grade 8 school. It does not net out the Grade 7 mark, so it records where a school’s students stand after the move, not how the school graded them. It correlates 0.966 with the mean Grade 8 mark of a school’s students, each measured from their Grade 8 school’s mean, and 0.420 with the baseline (Pearson, averaged across the 2018 and 2022 Gauteng Mathematics cohorts).

XGBoost residuals and Double ML (Methods 6–7) fit one flexible Grade 7 $\rightarrow$ Grade 8 curve to marks measured from each Grade 8 school’s mean, with no fixed effect for the Grade 7 school, and give each Grade 7 school its students’ mean residual; Double ML is the same fit, cross-fitted over five folds. Their correlations with the baseline are 0.710 and 0.709. The gap comes from the missing fixed effect, not from the flexible curve: a straight line fitted the same way gives mean residuals that correlate 0.773 with the baseline and 0.960 with the XGBoost residuals. Without the school’s own effect in the fit, the Grade 7 slope is estimated

¹Within a baseline-mark slice the bipartite Grade 7 to Grade 8 network thins relative to the full sample, and the `fixest::feols` package code correspondingly emits a warning about irregular fixed effects for every slice: the runtime warning that triggers when the bipartite graph has more than one connected component, so school fixed effects are identified only up to one constant per component. We computed the exact component structure on each slice’s edge list; the warning turns out to be largely cosmetic in our setting. Across the ten quintile slices in Table 24, each slice’s bipartite graph decomposes into between 9 and 17 components, but the largest component contains between 98.4% and 99.4% of the Grade 7 schools in the slice. The remaining one to two per cent of schools sit in small isolated sub-graphs (typically two- or three-school islands of feeder–destination pairs that no longer connect back to the main network once the slice is taken), and each contributes its own within-component normalisation. Since the dominant component captures essentially the whole network in every slice, the slice-level reliability ratios in Table 24 should be read as effectively unbounded claims for that dominant component; the within-component normalisations are a footnote on the headline number rather than a binding constraint on it. The pooled-sample numbers in §4.2 of the main text remain the headline reliability the paper relies on.

Figure B6: Within- vs. between-quintile variation in school effects

Each violin shows the distribution of feeder-school effects within a poverty quintile.

Gauteng, Mathematics. A high within-quintile share is the project analogue of CFR14's '85% within-school' res

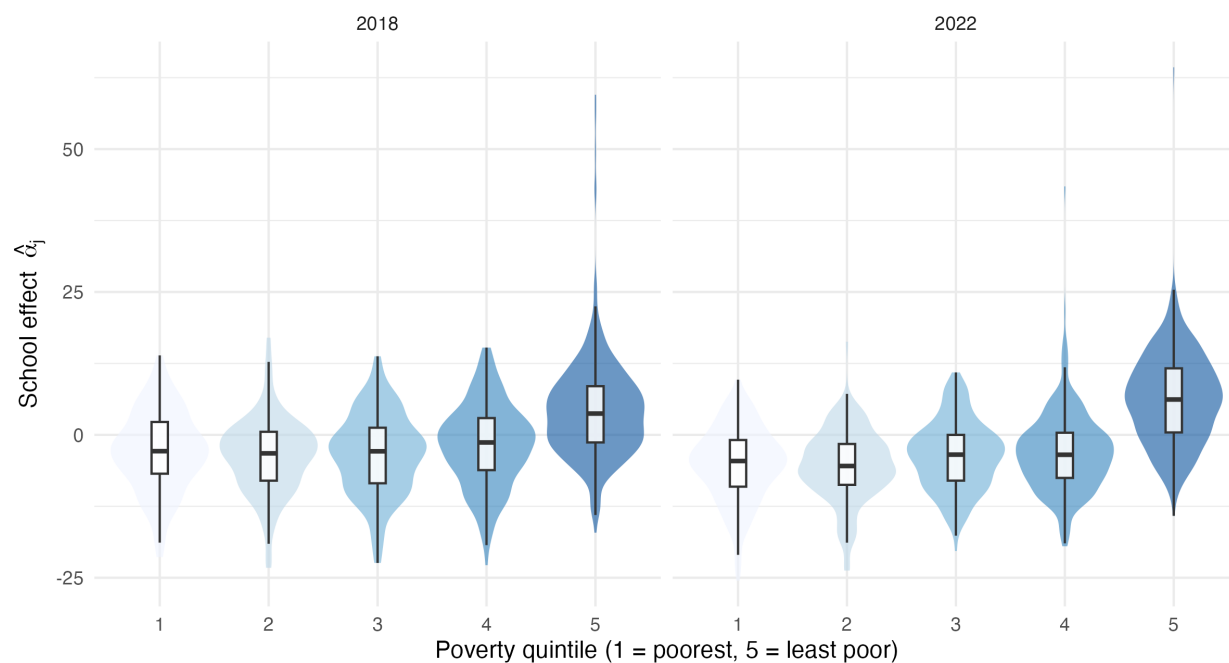


Figure 15: Within- vs between-partition share of the variance of $\hat{\alpha}_j$ across the two coarse partitions (national poverty quintile and district). *Sample:* Gauteng core, 2018 and 2022 cohorts.

partly across schools, so each school’s mean residual also carries part of its students’ Grade 7 standing relative to their Grade 8 classmates.

Table 25: Alternative estimators of the school effect: correlation with the baseline and dispersion (Gauteng Mathematics, 2018 and 2022 cohorts averaged).

Method	Description	Correlation with baseline	SD of effects (raw marks)	Median SE (raw marks)
Baseline TWFE	Linear, OLS	1.000	9.02	1.251
Quadratic TWFE	Add Gr7 ² as additional regressor	0.994	8.52	1.246
Piecewise linear	Hockey-stick spline at median Gr7 mark	0.998	8.68	1.249
EB shrinkage	Empirical Bayes shrinkage to cohort mean	0.998	8.63	—
Percentile rank equating	Within-class percentile ranks, mean by Gr7 school	0.420	0.06	—
XGBoost residuals	XGBoost on demeaned data; school = mean residual	0.710	4.58	—
Double ML	5-fold cross-fitted XGBoost; school = mean residual	0.709	4.59	—
Cross-classified HLM	Crossed random intercepts (lme4::lmer, REML=FALSE)	0.987	7.76	—

Each row’s school effects are correlated with the baseline two-way fixed-effects estimate (Pearson) in each cohort, and the two cohorts are averaged. Standard deviations and median standard errors are in raw Grade 8 mark points, except percentile rank equating’s, which are on a 0–1 rank scale; the baseline’s standard deviation is the same wedge that Table 3 of the main text reports in within-province SD. The first four rows and the last estimate the wedge; percentile rank equating, XGBoost residuals and Double ML estimate other objects (see the text above). Median standard errors are reported only where the method yields one. The missing-value glyph for this document is an em dash. Sample: Gauteng core, 2018 and 2022 cohorts.

For downstream applications using the calibrated learning measure as a mark-based school-level ranking, the parametric estimators (§4.2 of the main text) are the appropriate reference set. The rank divergence of percentile rank equating is not a failure of robustness but a consequence of measuring a conceptually different quantity; for applications that specifically require within-class rank comparisons, percentile rank equating may be preferred over the mark-point estimators.

H.7: Reliability of the calibrated measure: variance decomposition and split-sample reproducibility

The main text states the headline reliability ratios and split-sample correlations of the calibrated school effect; this table carries the full decomposition behind them. The variance-decomposition rows give the Kane–Staiger signal share of the cross-school variance in $\hat{\alpha}_j$ (raw Grade 8 mark points): the signal variance is the total variance of the estimated effects less the mean squared standard error, and the reliability ratio is signal over total. The split-sample rows re-estimate the calibration on two independent random halves of each school’s students and report the Pearson correlation between the two halves’ estimates across schools, with its Spearman–Brown projection to the full sample.

Table 26: **Most of the variance in the calibrated measure is signal, not noise.** *Variance decomposition and split-sample reproducibility (Gauteng Mathematics).* **Variance decomposition** rows 1–4: signal-over-total reliability (Kane–Staiger) applied to fixed-effects estimates $\hat{\alpha}_j$. Units: raw Grade 8 mark points. Row 3, $\text{Var}(\text{signal}) = \text{Row 1 minus Row 2}$ (floored at zero). Row 4, reliability ratio $\lambda = \text{Var}(\text{signal}) / \text{Var}(\hat{\alpha}_j)$. **Split-sample** rows: Pearson correlation between independent half-sample estimates and its Spearman–Brown projection to the full sample; median over ten random splits per cohort; no pooled equivalent (split not defined across cohorts). The cross-year correlation is reported in [Appendix G](#). **Sample:** Gauteng core, 2018 and 2022 cohorts.

Statistic	2018	2022	Pooled
Var(estimated effects)	83.86	78.79	81.25
Var(noise) = mean(SE ²)	2.552	2.289	2.418
Var(signal)	81.31	76.50	78.83
Reliability ratio	0.970	0.971	0.970
Split-half r	0.884	0.879	–
Spearman-Brown reliability	0.939	0.935	–

H.8: Is the wedge difficulty or differential persistence?

§4.5 of the main text reports that a one standard deviation difference in the no-fee indicator is associated with a wedge that is 0.49 cross-school standard deviations lower. Reading that gap as difficulty rests on the assumption that learning persists at a common rate across the Grade 7 to Grade 8 window (§3 of the main text). A no-fee school whose students score below their Grade-7-implied level on the shared Grade 8 anchor is booked as easier Grade 7 assessments. The same shortfall would arise, in the baseline design, if those students instead learned or retained less across the window, which spans the year-end holiday: if holiday learning loss is itself socioeconomically graded, it enters the wedge indistinguishably from easier assessments. Three re-estimations on the Gauteng Mathematics student panel separate the two readings.

The gradient survives both discriminating tests. It does not shrink when the Grade 7 pass-through is allowed to differ by poverty and by achievement, and it does not grow as the Grade 8 anchor moves later in the year. The third test shows the shortfall is a level shift plus a component that widens with Grade 7 achievement, which leaves a bounded part of the gap that the design does not attribute to marking alone.

Allowing the Grade 7 slope to differ by poverty and achievement. The wedge imposes a single Grade 7 pass-through on every school. Where the true pass-through is steeper for richer or higher-scoring schools, that misspecification loads a poverty-correlated term into the wedge and could manufacture the no-fee gradient. It is steeper in the richest schools. Estimated separately by poverty quintile, a Grade 7 mark one within-province standard deviation higher predicts between 0.46 and 0.56 of a standard deviation more on the Grade 8 mark in the four poorer quintiles and between 0.67 and 0.72 in the richest, against 0.60 and 0.59 under the single slope; [Appendix H.3](#) finds the same ordering in raw marks. Table 27 re-estimates the wedge under a single slope, a poverty-quintile-specific slope, and a school-mean-Grade-7-quintile-specific slope, and re-runs the no-fee association on each. The gradient is unmoved: across the eighteen specification-by-weighting cells it retains between 0.99 and 1.01 of its single-slope value. Imposing a common slope does not create the gradient.

Table 27: **The no-fee gradient survives poverty- and achievement-specific Grade 7 pass-through.** *Precision-weighted no-fee coefficient (headline atlas specification), in cross-school SD of the wedge, with the Grade 7 slope estimated three ways, and the fraction of the single-slope coefficient retained.* **Sample:** Gauteng Mathematics, 2,563 school-cohorts, both cohorts pooled. Associations only; no causal interpretation.

Grade 7 pass-through	No-fee coef. (SD)	Retained fraction
Single slope (baseline)	0.494	1.00
Poverty-quintile-specific	0.497	1.01

(continued)

Grade 7 pass-through	No-fee coef. (SD)	Retained fraction
Grade 7-quintile-specific (spline)	0.489	0.99

Moving the Grade 8 anchor later in the year. If the gap were ongoing differential fade, it would widen as the anchor moves from Term 2 to Term 4 and more of the transition window elapses. Table 28 re-estimates the wedge on the Grade 8 Term 2, Term 3, and Term 4 marks. The no-fee gap, in units of the wedge SD at each anchor, is flat across the horizon: 0.50, 0.48, and 0.51 at Terms 2, 3, and 4. The gap drifts by 0.0038 of its SD per term ($p = 0.036$), negligible next to the half-SD level it modifies. The gap is a stable offset already fixed by the Term 2 anchor. In common units it widens at Term 4 by a factor of 1.10, close to the ratio of the Term 4 to the Term 2 persistence slope (1.10 and 1.11), which is what difficulty alone predicts: a primary school’s offset enters each anchor scaled by that term’s persistence (Appendix D).

Table 28: **The no-fee gap is flat as the Grade 8 anchor moves later in the year.** *No-fee gap, in units of the wedge SD at each anchor, re-estimating the wedge on the Grade 8 Term 2, Term 3, and Term 4 marks. **Sample:** Gauteng Mathematics, 2,562 school-cohorts with all four marks, both cohorts pooled. Associations only; no causal interpretation.*

Grade 8 anchor term	No-fee gap (wedge SD)	SE
Term 2	0.500	0.024
Term 3	0.485	0.023
Term 4	0.507	0.024

Splitting a level shift from a slope difference. A pure difficulty offset shifts every no-fee student’s mark down by the same amount, whatever their Grade 7 score. Table 29 fits the Grade 8 z-score on the Grade 7 z-score, a no-fee indicator, and their interaction, within destination-by-cohort cells and clustered on origin school. The no-fee level shift is -0.274, close to the school-level 0.47 within-province SD gap. The interaction is not zero: the Grade-7-by-no-fee term is -0.149, about 0.27 of the fee-school pass-through, so the no-fee shortfall widens with Grade 7 achievement. A uniform offset is therefore refuted. The scale-dependence is jointly consistent with non-uniform, top-compressing marking, with noisier no-fee Grade 7 marks attenuating the within-group slope, and with genuinely larger learning differences among higher-scoring students; the first two are assessment channels and the third is a learning channel, and the design does not separate them. This bounded, higher-achiever component is why §4.1 of the main text states the wedge as mostly a difference in difficulty and does not attribute all of it to marking. The flatter slope follows poverty far more than difficulty. With primaries grouped into thirds by their wedge in the other cohort, the Grade 7 pass-through of primaries with harder and easier assessments differs by 0.116 overall, by -0.004 (standard error 0.013) among no-fee primaries and by 0.054 (0.016) among fee-charging ones, while the no-fee shortfall in pass-through is 0.170 before and 0.162 after each wedge third has its own slope. Letting the slope differ by wedge third, fee status or both barely moves the wedge: read at the province-mean mark, each refitted wedge has a Spearman correlation of at least 0.995 with the published one and a standard deviation of 0.47 to 0.48 within-province SD.

Table 29: **The no-fee shortfall is a level shift plus a component that widens with Grade 7 achievement.** *Student-level regression of the Grade 8 z-score on the Grade 7 z-score, a no-fee indicator, and their interaction, within destination-by-cohort cells; standard errors clustered on origin school. **Sample:** Gauteng Mathematics, 240,654 students, both cohorts pooled. Associations only; no causal interpretation.*

Coefficient	Estimate	SE	p-value
Grade 7 pass-through (fee schools)	0.550	0.017	< 0.001
No-fee level shift	-0.274	0.020	< 0.001

(continued)

Coefficient	Estimate	SE	p-value
Grade 7-by-no-fee interaction	-0.149	0.018	< 0.001

Appendix I: Comparison with Liu, Gershenson, and Anthenelli (2025)

Liu et al. (2025) develop a closely related approach to the same measurement problem in a US setting (Maryland and North Carolina middle and high schools), using transitions between schooling stages and a receiving-institution fixed effect to anchor sender-school grading practices onto a common scale. The two papers were developed independently and contemporaneously and converge on the same core design move: when students disperse from heterogeneous sending institutions into a receiving institution where assessment is common across formerly separated peers, the gap between input and output marks identifies the difference between the sending school’s grading scale and the common downstream scale, and a fixed effect on the receiving unit absorbs the latter. [Appendix G](#) benchmarks our cross-year correlation of 0.624 against their school-level range of 0.27–0.86. This appendix consolidates the broader methodological and interpretive contrast that the body does not address.

Table 30 summarises the principal differences we observe in design and framing across the two papers (the table is read as our authored comparison: where Liu et al.’s paper does not address a topic (e.g. measurement-error treatment of generated regressors, or validation against synthetic ground truth), the “Not addressed” / “None” entries record the contrast with our §5.2 of the main text and §5.1 of the main text work, not a claim Liu et al. themselves make).

Table 30: Methodological and framing differences between this paper and Liu et al. (2025)

Dimension	This paper	Liu et al. (2025)
Setting	Gauteng (extended to five additional provinces in Appendix D)	Maryland, North Carolina
Stage transitioned	Grade 7 → Grade 8 (primary → secondary)	Grade 8 → Grade 9 (middle → high), and Grade 5 → Grade 6 in MD
Subject coverage	Mathematics only	All subjects pooled into GPA
Receiving anchor	Grade 8 <i>grade-cell</i> fixed effect	Receiving <i>high-school</i> fixed effect
Sample restriction	Full Gr 7 → Gr 8 linked sample within the largest connected component	Algebra-I-in-Gr 9 students (about 60% of the cohort)
Estimator	Single-stage two-way fixed effects	Two-step value-added with leave-one-out school effects
Estimand framing	Scale-adjustment parameter (explicitly composite)	Value-added measure (with accountability applications)
Measurement-error treatment	Central (heaping documentation, structural de-heaping and MC evidence, downstream-rank stability: §5.2 of the main text)	Not addressed
Validation against synthetic ground truth	Monte Carlo against four threat scenarios (§5.1 of the main text; heaping reported in §5.2 of the main text)	None
External validation	Internal measurement properties only — no contemporaneous census standardised test exists	Predictive validity against Gr 9 GPA, retention, on-time graduation, and college enrolment
Use-case framing	Research input for cross-school comparison	State accountability indicator (ESSA “fifth indicator”)

The deepest difference is interpretive. Liu et al. (2025) frame the recovered effect in standard value-added language, measuring a school’s “contribution” to its students’ subsequent grades, and propose an accountability use case. We explicitly disavow any such application: §3 of the main text reframes $\hat{\alpha}_j$ as a scale-adjustment parameter that is observationally equivalent to a composite of student selection, instructional effectiveness, and peer and contextual effects (the receiving school’s assessment difficulty and grade-level instruction are absorbed by γ_k), and the use-case row of Table 30 records our measure as a research input, not an accountability indicator, precisely because the composite cannot be decomposed. The same estimator therefore serves different purposes in the two framings. Where Liu et al. (2025) require predictive validity against subsequent outcomes to defend the causal claim, we require only comparability and reliability diagnostics to defend the scale-adjustment claim. Neither framing is wrong; each is appropriate to its setting and use case, and the two papers are best read as complementary applications of a shared design rather than as competing claims about a single estimand.

The two papers’ validation strategies are accordingly complementary. Liu et al. (2025) validate the *predictive content* of their measure against subsequent outcomes (Grade 9 GPA, Grade 9 retention, on-time high-school graduation, on-time college enrolment) that they observe in state longitudinal data and that we cannot observe in our setting. We validate the *measurement properties* of the estimator: within-year signal-to-noise reliability and split-sample Spearman-Brown reliability (§4.2 of the main text), cross-method invariance across linear, quadratic, piecewise, empirical Bayes and cross-classified hierarchical specifications (§4.2 of the main text), measurement-error robustness of school-level rankings under heaping correction (§5.2 of the main text), Monte Carlo recovery of synthetic ground truth across sorting, limited-mobility, attrition, and threshold-pushing threats (§5.1 of the main text), and reliability under baseline-achievement subgroup splits (Appendix H.5, the closest direct analogue of Liu et al. (2025)’s bottom-tertile heterogeneity exercise). Neither paper could have run the other’s validations given its data constraints. The fact that the long-run predictive results in Liu et al. (2025) are largely null is itself instructive: it tempers expectations for any future student-to-matric crosswalk extension in our setting, and reinforces our scale-adjustment framing as the more defensible interpretation in data-poor settings where causal-VAM validation is impossible by construction.

A notable contrast emerges in subgroup findings across the two settings. Liu et al. (2025) find that in North Carolina (though not in Maryland), the per-SD effect on Grade 9 GPA of attending a high-VA middle school is more than twice as large for bottom-tertile baseline-math students as for top-tertile students. In our Grade 7 → Grade 8 mathematics setting, our subgroup analysis (Appendix H.5) examines the *dispersion* of recovered school effects, finding that it rises monotonically across the baseline-mark quintiles, with top-quintile dispersion roughly 50% larger than bottom-quintile dispersion in 2018 (and a similar but slightly compressed gradient in 2022; Table 24). Because the metrics differ (they examine mean effect sizes, while we examine the variance of the recovered school effects), the two results are not directly comparable. Nevertheless, the contrast shows how different segments of the achievement distribution can drive the identifying variation in different institutional environments.

Several Liu et al. (2025) design choices we deliberately do *not* adopt, in each case for principled reasons rooted in our institutional setting. First, single-subject mathematics rather than pooled GPA: Appendix C shows that pooling across subjects masks subject-specific signals because the South African school system does not apply a unified grading apparatus across subjects in the way a US high school does. Second, the full linked sample rather than a coursework-based subset: Liu et al. (2025) restrict to Algebra-I-in-Gr 9 students, excluding roughly 40% of the cohort; our estimand applies to the full population of mathematics-taking Grade 7 → Grade 8 movers within the connected component. Third, a Grade-8 *grade-cell* fixed effect rather than a receiving-school fixed effect: students in the same destination grade sit the same assessment, so the grade-cell effect absorbs destination assessment difficulty and grade-level instruction, a finer anchor than a whole-school effect. The source data carry no Grade 8 class identifier, so the anchor is the grade cell, not the class; residual across-class marking differences within a destination grade fall into the idiosyncratic error (§3.3 of the main text), and the “Grade 8 class” unit in the network descriptives (Table 2 of the main text, Panel C) is this destination grade cell. Fourth, a research-input framing rather than an accountability-indicator framing: the use-case mapping of Table 30 explicitly excludes accountability applications, while Liu et al. (2025)’s policy section explicitly proposes their measure as one.

Appendix J: Using These Estimates Downstream

This appendix is a short user guide for analysts who plan to use the calibrated learning levels as inputs to a second-stage analysis. The aim is to flag the inferential pitfalls that arise when a generated regressor (or outcome) is treated as if it were directly observed, and to point to the sanctioned corrections.

What is delivered. The estimation pipeline produces a school-by-cohort-year panel of point estimates $\hat{\alpha}_s$ and their associated standard errors $\widehat{\text{se}}(\hat{\alpha}_s)$, with one file per province and a pooled file across provinces. Both include the columns `baseline_emis` (the Grade 7 school identifier), `cohort_year`, `province`, `subject`, `estimate` ($\hat{\alpha}_s$, in within-province SD of Grade 8 marks), `se` ($\widehat{\text{se}}(\hat{\alpha}_s)$), `ci_lower/ci_upper`, `n_students`, `n_gr8_destinations`, and `quintile`. Each province and cohort’s wedges are measured from their own zero, so subtract the fit’s mean before comparing wedges across fits. A merged school-analysis panel adds school covariates. It also carries `comparable_score`, an older construction that adds each fit’s wedge to the raw Grade 7 mean and so mixes the two scales; use the calibrated level of §3.4 of the main text instead.

The two-step inference problem. When $\hat{\alpha}_s$ enters a downstream regression (as a regressor, e.g., a production-function specification with calibrated learning on the right-hand side, or as the outcome, e.g., a regression of calibrated learning on school resources), the first-stage sampling uncertainty must be propagated to the second stage. Standard errors computed from the second-stage regression alone will be too tight, treating $\hat{\alpha}_s$ as if it were the population parameter rather than an estimate. The magnitude of the distortion scales with the ratio of estimation noise to true cross-school variance: the reliability decomposition in Table 26 implies that in Gauteng this ratio is small, but the correction is not negligible and is larger in provinces with sparser mixing networks (Table 13).

Recommended approaches. Three corrections, in rough order of preference for typical applications:

1. **Paired bootstrap resampling both stages (default).** Resample Grade 7 schools (or, for the first stage, student-class clusters) with replacement, re-estimate $\hat{\alpha}_s$ and $\widehat{\text{se}}(\hat{\alpha}_s)$ from scratch on each bootstrap draw, then re-run the second-stage regression. The bootstrap standard errors of the second-stage coefficients automatically incorporate both sources of uncertainty. This is the safest default because it makes no parametric assumption about the form of the first-stage sampling distribution.
2. **Errors-in-variables correction (Fuller) when $\hat{\alpha}_s$ is a regressor.** If the second-stage model is a linear regression with $\hat{\alpha}_s$ on the right-hand side, a classical measurement-error correction (see Fuller 1987) using the school-specific $\widehat{\text{se}}(\hat{\alpha}_s)^2$ as the known error variance recovers consistent coefficients. This is lighter-weight than the bootstrap and sufficient when the measurement error can be treated as classical.
3. **Precision-weighted regression when $\hat{\alpha}_s$ is the outcome.** If the second-stage model is a regression of $\hat{\alpha}_s$ on school characteristics, weight observations by $1/\widehat{\text{se}}(\hat{\alpha}_s)^2$. This does not by itself correct inference, but it correctly downweights noisily estimated schools, and combined with bootstrap standard errors it recovers a well-behaved estimator. An empirical-Bayes shrunk version of $\hat{\alpha}_s$ is also available (see [Appendix H.1](#)) and can be substituted when the application tolerates the shrinkage-induced bias in exchange for lower variance.

What not to do. Two misuses are common enough to flag explicitly. First, the calibrated learning level is a *school-level* parameter and should not be interpreted at the student level: a student attending a school with $\hat{\alpha}_s > 0$ is not guaranteed to be a high-achieving student, and individual student inference from $\hat{\alpha}_s$ is not supported by the estimation framework (this is the concern raised in [Appendix Q](#)). Second, $\hat{\alpha}_s$ is not inter-temporally comparable without additional assumptions: the 2018 and 2022 estimates are each normalised to their own within-province cohort and the scale is not preserved across cohorts (see [Appendix G](#)). Panel specifications that pool the two years, or that difference $\hat{\alpha}_s$ across cohorts, should not treat the year-on-year change as an estimated quantity without a separate linking argument. When in doubt, analyse each cohort as its own cross-section.

Appendix K: External Validation Against NSC Mathematics Results

This appendix tests whether Grade 8 school-level calibrated means are more closely rank-correlated with an external school-quality benchmark than raw means. The benchmark is school-level National Senior Certificate (NSC) Mathematics results, matched via school EMIS code. NSC Mathematics (subject code 57) measures end-of-schooling attainment under a nationally standardised examination, making it a plausible long-run external criterion for the school-quality signal embedded in our estimator.

Year alignment. The 2018 cohort is aligned to NSC 2022 on a cohort-tracking basis: Grade 8 students in 2018 would sit the NSC approximately four years later. The 2022 cohort is aligned contemporaneously to NSC 2022 because NSC 2026 data are not yet available. Both cohort years therefore use the same NSC 2022 school-level means.

Results. Figure 16 shows rank scatter plots for raw and calibrated Grade 8 school means against NSC school means, separately by cohort year. The Spearman rank correlation is $\rho_{\text{raw}} = 0.234$ for the 2018 cohort and $\rho_{\text{raw}} = 0.261$ for the 2022 cohort; calibration yields $\rho_{\text{cal}} = 0.299$ and $\rho_{\text{cal}} = 0.330$ respectively, a lift of 0.065 and 0.069. Both correlations are statistically significant at conventional levels ($p < 10^{-60}$ throughout). The match rate exceeds 97% of Grade 8 schools in both cohorts.

Interpretation. Calibration raises the correlation, but it stays modest. A de-identified national extract of the same administrative records, which links each Grade 8 student to that student’s own NSC result, shows which of the candidate explanations bind. The extract points at the 2025 examination, matches students rather than schools and carries about 5% fewer students per school than ours, so its correlations are read against the figures above for where they fall, not as the same estimate.

First, the NSC and the Grade 8 mark are four years of schooling apart, and the link between a school-set mark and the examination weakens steadily with that distance. Regressing the NSC Mathematics score on the school-set Mathematics mark with a fixed effect for the school that set the mark, the slope falls from 0.967 at Grade 12 to 0.551 at Grade 8, and the within-school R^2 from 0.883 to 0.223. The Grade 12 figure overstates agreement between school and examiner, because a quarter of the final NSC score is the school’s own assessment.

Second, student mobility between Grade 8 and matric is not what holds the correlation down. Setting each Grade 8 school against the mean NSC score of its own students, wherever they wrote the examination, gives a Spearman correlation of 0.256 over 3,348 schools. With the school means standardised, as they are in those figures, it is 0.279, above both. The students found again in the examination are a selected group: 16.5% of Grade 8 Mathematics students carry an NSC Mathematics mark, and they score 1.39 standard deviations above the rest on their Grade 8 mark.

Third, a single year’s NSC school mean is itself a noisy benchmark, and nothing in either data set bounds that noise.

Adjusting the Grade 8 intake in other ways moves the correlation little. The closest analogue on the student-linked design subtracts each Grade 8 school’s destination effect, estimated on the following cohort and matched to the Term 4 mark; it raises the correlation from 0.237 to 0.276 on the same 2,242 schools, a smaller rise than calibration gives the school-level figures above. Subtracting an intake-school effect estimated from students who change secondary school lowers it from 0.311 to 0.161 on 2,820 schools, even with the effect estimated on a split sample so that it never sees the outcome: with a handful of movers per school, the subtraction removes more signal than difficulty. That destination effect does not follow how the same school marks its own Grade 12 students. Against the gap between a secondary school’s Grade 12 Term 4 mark and its students’ NSC marks, pooled over NSC 2021, 2022 and 2024, the destination effect estimated for the 2021 Grade 7 cohort at Grade 8 Term 1 correlates at 0.044 over 1,399 schools, and at 0.117 when estimated at Term 4, although it correlates at 0.424 with itself across cohorts and a school’s own Grade 8 marks set against the NSC correlate with its own Grade 12 marks at 0.239 (precision-weighted correlations, within province). The Grade 12 mark in the extract is the school’s own, before Umalusi’s moderation. Gauteng is absent from these data, so this tests the replication’s destination effect in the other six provinces, not $\hat{\gamma}_k$.

The Grade 8 school effects are more stable across cohorts than this benchmark can confirm: the Spearman

Gr8 vs NSC school means: raw vs endline_emis-calibrated (Mathematics)

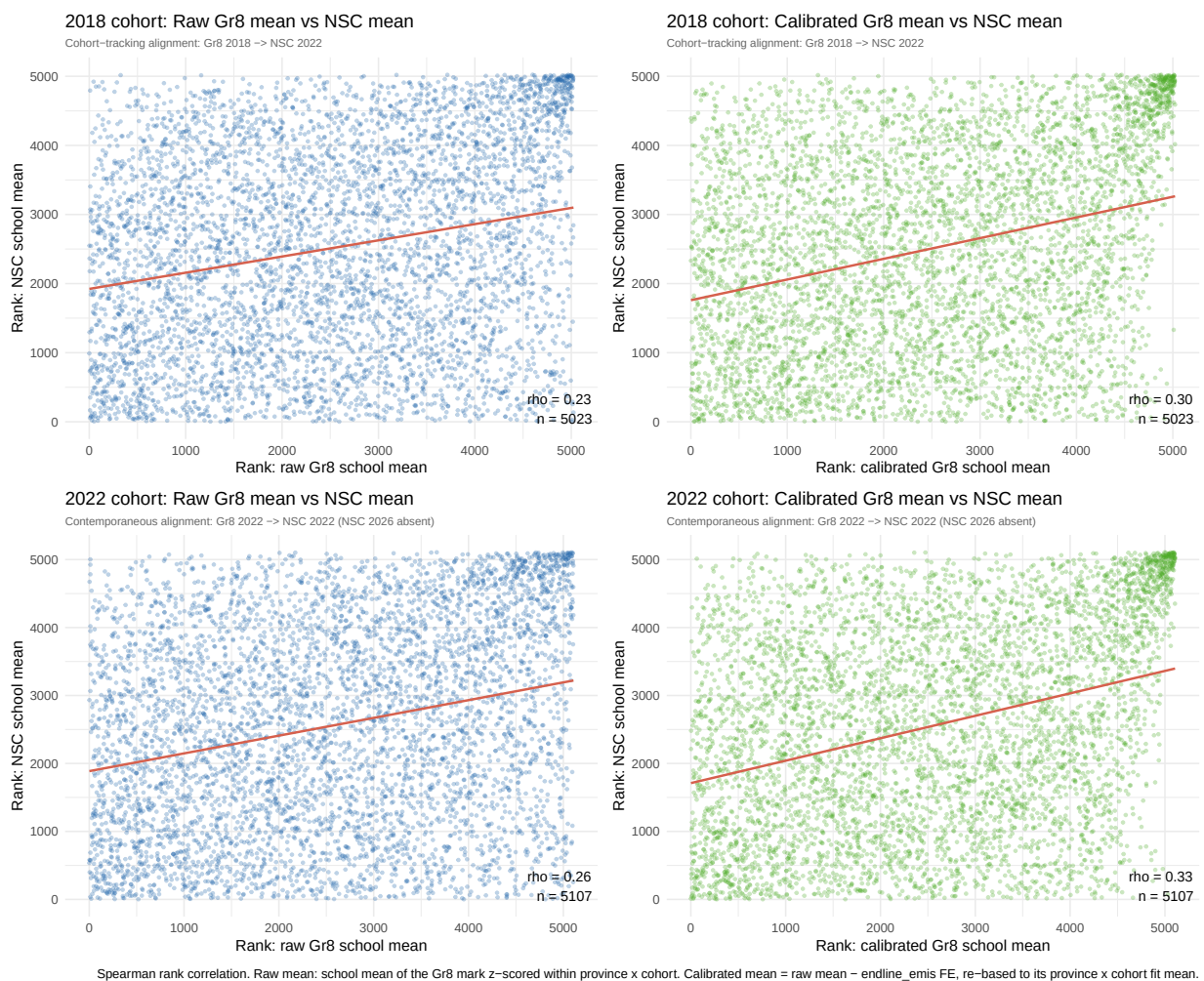


Figure 16: Rank scatter plots: Grade 8 school means (raw and calibrated) against NSC 2022 Mathematics school means. Each point is a secondary school. Left column: raw Grade 8 mean, in within-province standard deviations. Right column: calibrated mean, the raw mean minus the school's endline fixed effect, each effect taken relative to the mean school of its province and cohort. Rows correspond to the 2018 and 2022 cohorts. Spearman ρ and sample size are annotated in each panel. *Sample*: secondary schools of the six-province Mathematics panel matched to an NSC 2022 Mathematics school mean, by cohort; the number of schools is annotated in each panel.

correlation between the 2018 and 2022 endline estimates is 0.372 over 6,075 schools, above every NSC rank correlation. These results are consistent with the measurement-properties framing developed in §4.2 of the main text and [Appendix I](#), and with the cautionary note in Liu et al. (2025) that long-run predictive validity of school-quality measures is generally modest even in data-rich settings.

What a single school fixed effect on the examination recovers

Figure 17 reports, cell by cell, the correlation across schools between a single school fixed effect, fitted on students who stay in the same school, and each part of a two-way fit with separate fixed effects for the intake school and the exit school: the intake-school effect, the exit-school effect and their sum. The single fixed effect follows the sum and neither component. Across 45 cells, each a grade, a subject and either a province or all provinces pooled, its correlation with the sum runs from 0.9944 to 0.9999, and its correlation with the exit-school effect alone from 0.015 to 0.607. The outcome in every cell is the NSC score, the final blended mark of school-based assessment and an external examination. The exit-school effect is not offered as an ordering of schools. Table 31 reports its split-half reliability, Spearman-Brown corrected, at 0.377 for Mathematics at the Grade 8 intake, against 0.939 for the wedge in the Gauteng-core Grade 7 2018 Mathematics cohort. The cells span intake marks from Grades 8, 9 and 10 and the subjects Mathematics and English First Additional Language, and the rows labelled all provinces average the province rows, weighting each by its schools. The two-way fit separates the intake-school effect from the exit-school effect through students who change school between the intake grade and the examination; the single fixed effect is fitted on students who do not.

Table 31: **The exit-school effect reproduces far less well than the paper’s wedge.**

	Reproducibility			Dispersion		
	Schools (1)	Split-half (2)	Reliability (3)	SD (4)	SE (5)	Interval (6)
<i>Panel A. This extract, the exit-school effect at Grade 8</i>						
Mathematics	1,320	0.232	0.377	0.683	0.015	[0.759, 0.819]
English FAL	2,701	0.389	0.560	0.491	0.010	[0.558, 0.594]
<i>Panel B. The paper, its wedge in Gauteng</i>						
Mathematics, 2018 cohort	1,414	0.884	0.939	—	—	—

Note. English FAL in the stub is English First Additional Language. Column (1) is the number of schools on which the two halves are compared; column (2) the correlation between the two half-sample estimates of the same school effect; column (3) that correlation projected to the full sample by the Spearman-Brown formula. Column (4) is the standard deviation of the estimated effect across schools, column (5) its cluster bootstrap standard error over 100 replications, and column (6) the 2.5th to 97.5th percentile of that bootstrap, which is the interval named in its head. Panel A reports the exit-school effect: the school-level fixed effect on the National Senior Certificate score, net of an anchored Grade 8 intake mark for the same pupils, in within-province standard deviations of that score. The National Senior Certificate score is the final blended mark, 25% school-based assessment plus 75% external. The exit-school effect is a different object from the wedge of Panel B, which is the Grade 7 origin school’s estimated assessment difficulty in within-province standard deviations of the Grade 8 mark. This table is the limit that travels with the single-fixed-effect result, and it is the reason no school ranking is offered on the exit-school effect: its Spearman-Brown reliability is 0.377 in Mathematics over 1,320 schools, against 0.939 for the paper’s own wedge over 1,414 schools. An effect that reproduces at that rate identifies a dispersion, not a position in a league table. The split half is learners split at random within intake school, each half re-fitted two-way on its own connected set; schools compared where both halves carry at least 10 learners. The bootstrap is a non-parametric cluster bootstrap, clusters = intake schools, drawn with replacement within province and re-keyed so a school drawn twice contributes two independent fixed effects; each replication is a full seven-province refit. Levels are not comparable between the two panels: the paper’s Gauteng-core sample carries roughly 5% more learners per

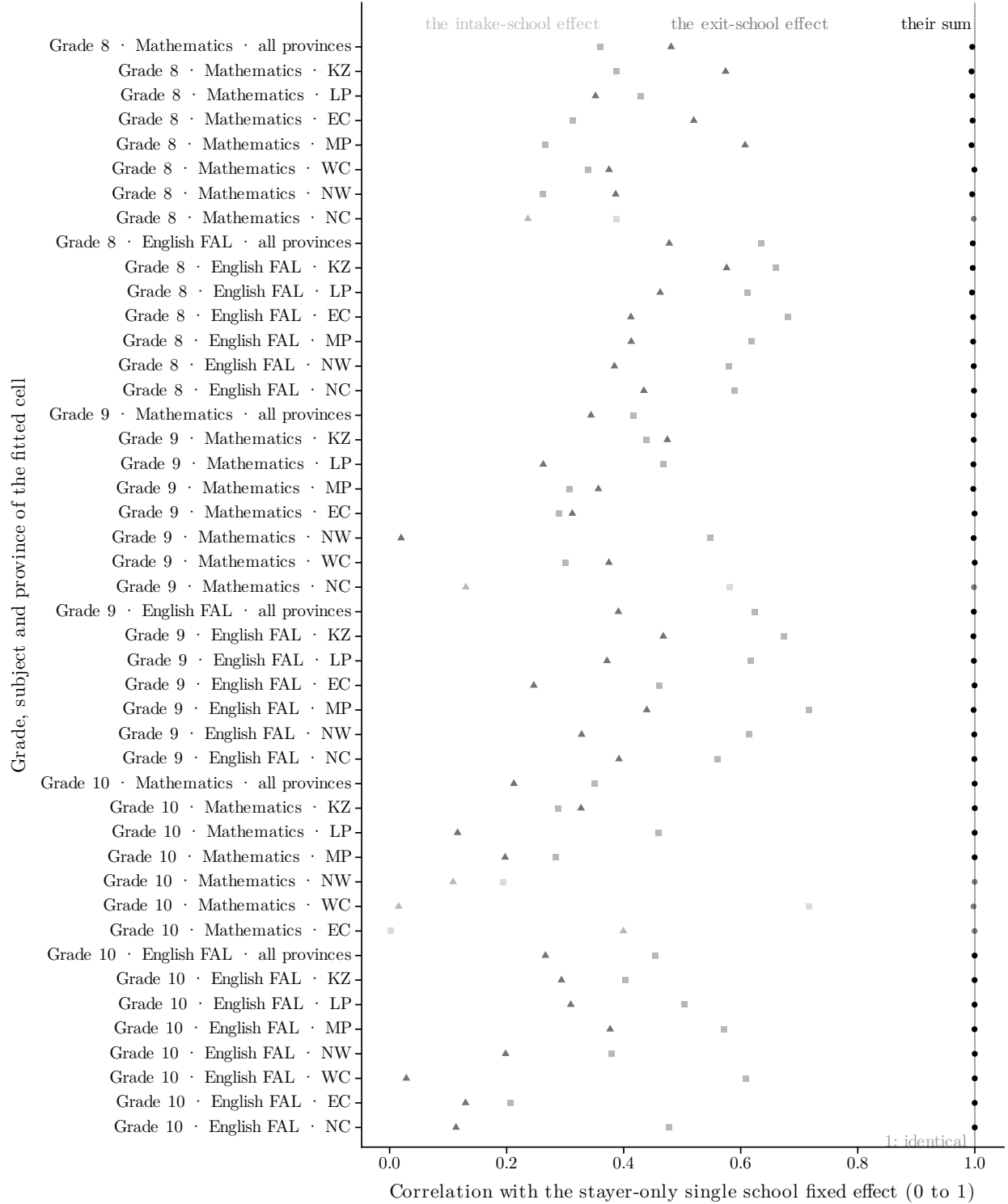


Figure 17: **A single school fixed effect recovers the sum, not either component.**

Note. Each row is one fitted cell: a grade, a subject and either a province or the pooled national figure. The horizontal axis is the correlation between the school fixed effect from a single-fixed-effect fit on stayers alone and, in turn, the sum of the two-way intake-school and exit-school effects, the exit-school effect by itself, and the intake-school effect by itself. The exit-school effect is the school-level fixed effect on the National Senior Certificate score net of an anchored intake mark; the intake-school effect is the fixed effect on the school that set the intake mark, the sign of the paper's wedge. The National Senior Certificate score is the final blended mark, 25% school-based assessment plus 75% external. Neither effect is a causal value-added estimate. Across the 45 cells the correlation with the sum runs from 0.9944 to 0.9999, against 0.015 to 0.607 for the exit-school effect alone and 0.002 to 0.717 for the intake-school effect alone. The grey vertical line at 1 is the benchmark, drawn behind the data. A cell drawn at half opacity compares fewer than 100 schools. **The exit-school effect is not offered here as a school ranking: its split-half reliability is 0.377 over**

school than this extract does, so only the shape and the slope of the comparison carry information. No cell of this table carries a significance star, because the source reports no p-value for a split-half correlation or for a bootstrapped dispersion; where stars appear elsewhere in this document they follow one ladder, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. A cell carrying the missing-value glyph is a quantity the source does not report: the paper’s split-half file carries no dispersion of the wedge and no bootstrap of it, so columns (4) to (6) of Panel B are empty in the source rather than empty by construction. The missing-value glyph for this document is an em dash. Source: analyses/sba_nsc_link/outputs/tables/tab5_design_b.csv panel C and outputs/tables/mathematics/tab10_split_sample_reliability_gauteng.csv, via outputs/tables/mathematics/tab_nsc_exit_power.csv, which records the source table and selector behind every cell. **Sample:** Panel A, pupils with a valid Grade 8 Term 4 2021 school-set mark at a school of at least ten such pupils and a same-subject score at their own on-time National Senior Certificate in 2025, seven provinces, split at random within intake school; 1,320 Mathematics schools and 2,701 English First Additional Language schools carry both halves. Panel B, the paper’s Gauteng-core Mathematics panel of 1,414 schools in the 2018 cohort.

The wedge against the student’s own NSC mark

Table 32 reports in full the specifications drawn in Figure 9 of the main text: the mark regressions of Panel A, the participation models of Panel B, standard errors clustered at the origin school, and the ratio of the wedge coefficient to the own-mark coefficient against its structural prediction, 1.91. Table 33 reports the wedge coefficient by province of the secondary school of entry, with and without the origin school’s mean mark, each against that province’s own prediction. The wedge is positive and significant in every province cell of both subjects; the North West Mathematics cell is the smallest and the weakest, and nothing is read into it in either direction. The object in both tables is the conditional increment of §4.7 of the main text, a different object from the school-level rank correlation with which this appendix opens; the two are reconciled there.

Two limits bind the mark regressions of Table 32. One is selection. They are estimated on the students who sat, 0.396 of the cohort, a sample selected on the wedge itself, since the wedge predicts sitting. Selection of that kind would be expected to pull the mark coefficients towards zero, and no bound on it is available here. The panel carries no assigned treatment for Lee bounds and no exclusion restriction for a selection model, and reweighting on observables assumes that sitting is ignorable given them, which is what the selection argument denies. The other is the blend. The outcome is the blended mark, whose school-based quarter is set by the Grade 12 school, a different school from the origin whose wedge is the regressor and one the fixed effect for the secondary school of entry absorbs, so the channel is bounded rather than removed.

Where the ratio shortfall comes from. Besides the part of a primary school’s assessment difficulty that persists, the 2020 wedge carries estimation error, a share of the Grade 8 marks of the same students whose NSC outcomes it predicts, and difficulty specific to the 2020 cohort. Each can be removed in turn on one sample, the 336,847 students whose primary school also has a wedge in the 2018 and 2022 cohorts. Removing estimation error, measured from the calibration’s standard errors, lifts the participation ratio from 1.74 to 1.86. Measured instead from the disagreement between wedges estimated on independent thirds of each primary school’s students, the error is 1.55 times as large, and the corrected ratio reaches 1.94. Re-estimating the wedge without each student’s own third works the other way, because a wedge built partly from a student’s own Grade 8 mark shares that mark’s information about the student’s later results. A split-sample instrument that removes both leaves the participation ratio at 1.85 (0.018), and instrumenting with the 2018 and 2022 wedges, which removes the 2020-specific part as well, gives 1.95 (0.025) against the benchmark of 1.91. The Mathematics mark runs from 1.47 through 1.57 to 1.30 (0.051), and the Mathematical Literacy mark from 1.65 through 1.77 to 1.74: within the 2020 cohort the two corrections together close less than half of either mark’s gap, and the part that persists across cohorts carries no more weight than the 2020 wedge. Sitting Mathematical Literacy, which overshoots the benchmark, overshoots by more after each correction, reaching 3.56 on the persistent part. Standard errors, in parentheses, come from a cluster bootstrap over origin schools that holds the calibration fixed.

The paper’s own 2018 cohort cannot be followed to the NSC student by student, because the term marks that link the two record systems begin in 2021. At school level it can: the mean Grade 7 2018 wedge of a secondary

school's intake predicts that school's mean NSC 2023 mark at 0.492 (0.030) across 3,142 secondary schools in Mathematics and 0.612 (0.029) across 3,400 in Mathematical Literacy, in national NSC-2023 standard deviations of the school mean per unit of the intake's mean wedge, conditional on the intake's mean raw mark and a province fixed effect, positive in every province, with robust standard errors in parentheses. That is an association between two school aggregates, on the part of the cohort's secondary schools that the school crosswalk reaches, with no within-school comparison and no student followed to the examination; it is a weaker object than the student-level test and does not relieve the cohort caveat on that test.

Table 32: **Students from origin schools with harder assessments sit the NSC more often and score higher.**

	Coefficient on			Ratio	Within R ²	Observations
	Wedge	Grade 7 mark	Calibrated score			
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Panel A. NSC 2025 mark, students who sat</i>						
Mathematics, M0		0.338 (0.009)			0.152	70,790
Mathematics, M4			0.457 (0.005)		0.191	70,790
Mathematics, M1	0.697*** (0.016)	0.458 (0.005)		1.524	0.195	70,790
Mathematics, M3	0.401*** (0.015)	0.532 (0.005)			0.208	70,790
Mathematical Literacy, M0		0.309 (0.007)			0.160	85,454
Mathematical Literacy, M4			0.469 (0.005)		0.204	85,454
Mathematical Literacy, M1	0.751*** (0.013)	0.463 (0.005)		1.620	0.207	85,454
Mathematical Literacy, M3	0.454*** (0.016)	0.516 (0.005)			0.214	85,454
<i>Panel B. Sitting the NSC, the whole entering cohort</i>						
Sat the NSC	0.340 (0.003)	0.196 (0.001)		1.737	—	394,353
Sat Mathematics	0.286 (0.003)	0.176 (0.001)		1.629	—	394,353
Sat Mathematical Literacy	0.054 (0.003)	0.020 (0.001)		2.678	—	394,353
Mathematics, given sitting	0.365 (0.005)	0.228 (0.002)		1.602	—	156,244

Note. Each row is one fitted specification; the coefficient sits on the line above its standard error. The outcome in Panel A is the student's NSC 2025 mark, standardised nationally over every 2025 sitter of the subject. The NSC mark is the final blended mark: 25% school-based assessment plus 75% external. The wedge is the Grade 7 origin school's estimated assessment difficulty, in within-province standard deviations of the Grade 8 mark, read from analyses/hs_value_added and never refitted; the own Grade 7 mark is the student's Grade 7 Term 4 mark, standardised within province and subject on the calibration sample, so it is in standard deviations of the Grade 7 mark, and its standard deviation in this sample is not exactly one. Coefficients are therefore in national NSC standard deviations per unit of the regressor. Every specification absorbs a fixed effect for the secondary school of entry, which absorbs a province fixed effect exactly because every secondary school sits in one province, and controls for sex, home language, population group, over-age

status, and the origin school's quintile, no-fee status and rural status, each carrying missingness as its own level. M0 holds the own Grade 7 mark and the controls; M1 adds the wedge; M3 adds the origin school's own mean Grade 7 mark as well. M4 is not a step of that ladder. It replaces the own Grade 7 mark with the calibrated score, the own mark and the wedge combined at the weights the calibration fixes and standardised within province and subject on the calibration sample, as the own mark is; column (3) is its coefficient. M1 lets the data choose the weights that M4 fixes. M0 and M4 do not nest each other, so comparing their within-R-squared compares fit and is not a formal test. Column (4) is the ratio of the wedge coefficient to the mark coefficient in M1, which the structural model of the module's specification predicts equals the reciprocal of the calibration slope, averaged over the six provinces, 1.912. The module also reports a naive comparison of the wedge coefficient against the calibration slope itself; that comparison is not dimensionally coherent and is not the test. Column (5) is the within-R-squared of each Panel A fit: the share of the variance of the NSC mark within secondary school of entry that the specification's regressors explain. Panel B fits the same right-hand side as a linear probability model on the whole entering cohort, whose outcomes are sitting the NSC at all (base rate 0.396), sitting Mathematics (0.180), sitting Mathematical Literacy (0.217), and sitting Mathematics conditional on sitting at all (0.453). It is the module's headline margin because it is the one estimated on a sample not selected on the regressor. Standard errors in parentheses are clustered at the Grade 7 origin school, the level at which the wedge varies. Stars follow the document ladder, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and mark the wedge coefficient, the one coefficient for which the module reports a p-value; an absent star is not a statement about significance. A blank cell is structurally empty: M0 contains no wedge, M4 contains neither the wedge nor the own mark, only M4 contains the calibrated score, and the ratio is defined on M1 alone. The missing-value glyph for this document is an em dash. It marks the within-R-squared of the Panel B models, which the module does not report. Source: analyses/alpha_nsc_validation/outputs/tables/tabA2_ladder.csv, tabA2_participation.csv and tabA2_ratios.csv, via outputs/tables/mathematics/tab_alpha_nsc.csv. **Sample:** the 394,353 students of the Grade 7 2020 cohort who entered a secondary school in one of six provinces (Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West) carrying a calibrated intake mark. Panel A is the 70,790 of them who sat NSC Mathematics in 2025, at 4,271 secondary schools of entry and 9,453 Grade 7 origin schools, and the 85,454 who sat Mathematical Literacy, at 4,299 and 9,708. Panel B is the whole cohort.

Table 33: The wedge predicts the NSC mark in every province and both subjects.

	Wedge				
	M1	M3	Ratio	Prediction	Observations
	(1)	(2)	(3)	(4)	(5)
<i>Panel A. Mathematics</i>					
EC	0.730*** (0.027)	0.438*** (0.029)	1.639	1.901	16,010
GT	0.689*** (0.054)	0.406*** (0.048)	1.291	1.815	7,672
KZN	0.739*** (0.034)	0.440*** (0.036)	1.643	2.113	14,884
LP	0.645*** (0.030)	0.373*** (0.031)	1.395	1.988	17,324
MP	0.739*** (0.045)	0.337*** (0.044)	1.618	1.758	10,325
NW	0.473*** (0.048)	0.297*** (0.051)	1.159	1.896	4,575
<i>Panel B. Mathematical Literacy</i>					
EC	0.749*** (0.025)	0.431*** (0.031)	1.637	1.901	19,440
GT	0.810*** (0.029)	0.580*** (0.041)	1.551	1.815	12,794

(continued)

	(1)	(2)	(3)	(4)	(5)
KZN	0.784*** (0.032)	0.467*** (0.039)	1.796	2.113	19,906
LP	0.692*** (0.037)	0.446*** (0.040)	1.658	1.988	13,978
MP	0.679*** (0.031)	0.371*** (0.051)	1.425	1.758	11,179
NW	0.746*** (0.039)	0.442*** (0.050)	1.470	1.896	8,157

Note. Each row is one fitted specification; the coefficient sits on the line above its standard error. The outcome in Panel A is the student’s NSC 2025 mark, standardised nationally over every 2025 sitter of the subject. The NSC mark is the final blended mark: 25% school-based assessment plus 75% external. The wedge is the Grade 7 origin school’s estimated assessment difficulty, in within-province standard deviations of the Grade 8 mark, read from analyses/hs_value_added and never refitted; the own Grade 7 mark is the student’s Grade 7 Term 4 mark, standardised within province and subject on the calibration sample, so it is in standard deviations of the Grade 7 mark, and its standard deviation in this sample is not exactly one. Coefficients are therefore in national NSC standard deviations per unit of the regressor. Every specification absorbs a fixed effect for the secondary school of entry, which absorbs a province fixed effect exactly because every secondary school sits in one province, and controls for sex, home language, population group, over-age status, and the origin school’s quintile, no-fee status and rural status, each carrying missingness as its own level. M1 adds the wedge to the own Grade 7 mark and the controls; M3 adds the origin school’s own mean Grade 7 mark as well. Columns (1) and (2) give the wedge coefficient in M1 and in M3. Column (3) is the M1 ratio of the wedge coefficient to the mark coefficient and column (4) the province’s own reciprocal calibration slope, which is what the structural model predicts column (3) should equal. Observations are the M1 cell count, which M3 shares. The North West Mathematics cell is the smallest of the exercise and the module’s specification marks it the one weak result; nothing is read into it, in either direction, without a standard error that separates it from the other five. Standard errors in parentheses are clustered at the Grade 7 origin school, the level at which the wedge varies. Stars follow the document ladder, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and mark the wedge coefficient, the one coefficient for which the module reports a p-value; an absent star is not a statement about significance. A blank cell is structurally empty: the ratio, the prediction and the observation count belong to the cell rather than to the standard-error line beneath it. The missing-value glyph for this document is an em dash. None arises here. Source: analyses/alpha_nsc_validation/outputs/tables/tabA2_ladder.csv and tabA2_ratios.csv, via outputs/tables/mathematics/tab_alpha_nsc.csv. **Sample:** the same Grade 7 2020 cohort as the main table, split by the province of the secondary school of entry. Cell sizes run from 4,575 to 19,906 students.

Appendix L: Monte Carlo bounding sweeps

The four headline Monte Carlo blocks in §5.1 of the main text condition on a single set of design choices: a chosen endogenous-mobility intensity ψ , a linear-in-means peer-effect specification, a Grade-7-only application of threshold-pushing, and an empirical attrition severity. This appendix collects the dedicated bounding sweeps that pin each of those choices against an empirical anchor and a published bias threshold, and identifies where the empirical signature lands relative to the threshold.

L.1: Endogenous-mobility bounding against the Andrews/KSS 5% threshold

The endogenous-mobility grid in §5.1 of the main text conditions on a chosen ψ and does not say what ψ the data actually correspond to. To anchor it, we compute on the empirical panel (Mathematics, 2018 and 2022 cohorts) the within-feeder Pearson correlation between the student’s Grade 7 baseline mark (demeaned

within feeder) and the estimated Grade 8 destination fixed effect $\hat{\gamma}$. This proxy is computable in both the data and the MC because it relies only on observables and on the same TWFE fit. Pooled across 22,148 within-feeder cells in 12 subject-province-cohort strata, the empirical signature is 0.021 (stratum means from -0.01 to 0.06). At observed network density (Figure 18), this signature maps to $\psi \approx 0.05$, at which the endogenous-mobility-induced shift in $V(\hat{\alpha})/V(\alpha)$ is 2.3 percentage points. The Kline et al. (2020) 5%-bias threshold requires $\psi^* \approx 0.119$, corresponding to an empirical signature of 0.048, which is 2.28 times larger than what we observe in the live data. The remaining 7.1% finite-sample inflation at $\psi = 0$ is sparsity-driven sampling noise (Figure 18), not endogenous-mobility bias, and shrinks towards zero as school size grows. On the data, the leave-out correction of Kline et al. (2020) measures that sampling noise directly: 5.4% of the estimated variance of the wedge in 2018 and 5.5% in 2022 (Appendix S.3). Endogenous mobility in the South African feeder network therefore sits well below the 5 per cent line; the level-side concern that the endogenous-mobility grid of §5.1 of the main text documents for $\psi \geq 1$ is a stress test, not a description of the empirical setting.

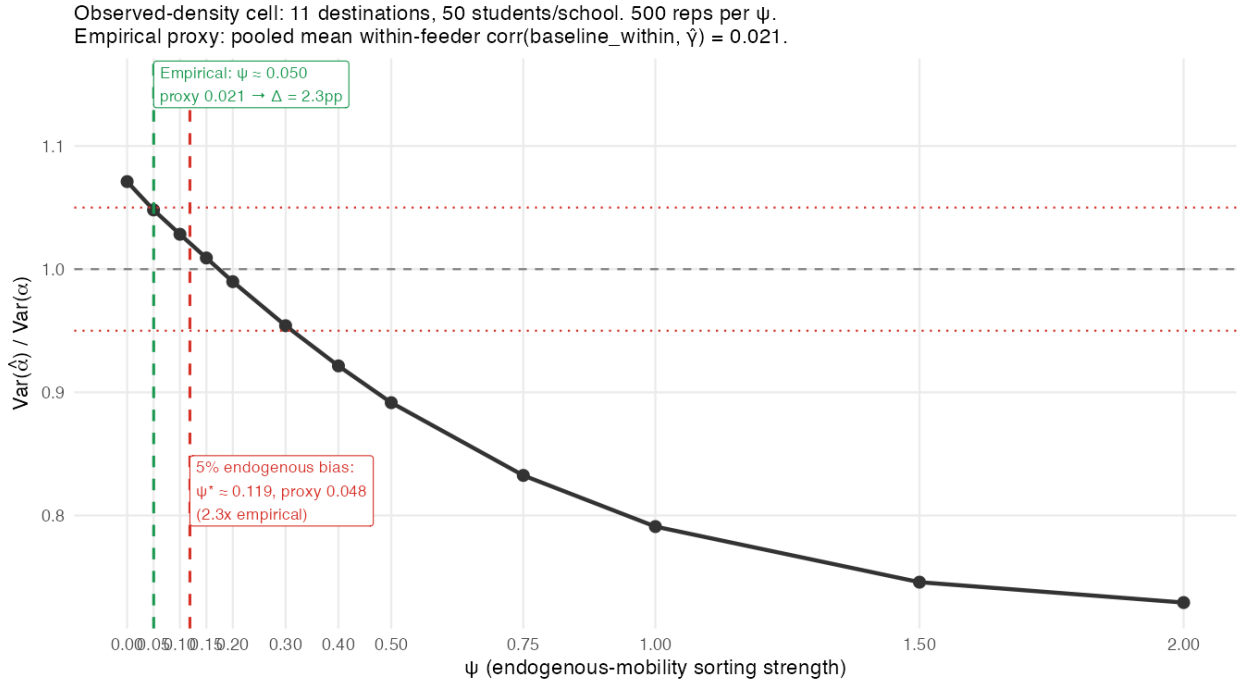


Figure 18: Bounding endogenous-mobility strength against the Kline et al. (2020) 5% threshold. Black curve: mean $\text{Var}(\hat{\alpha})/\text{Var}(\alpha)$ across 500 MC replications per ψ at observed network density (11 destinations, 50 students/feeder). Dotted red lines: $\pm 5\%$ deviation band. Green dashed line: ψ implied by the empirically-observed within-feeder $\text{corr}(\text{baseline}_{\text{within}}, \hat{\gamma})$ of 0.021. Red dashed line: ψ^* at which the endogenous-mobility-induced variance shift first exceeds 5 percentage points. Inverted using the matching MC-side proxy so the comparison is like-for-like. *Sample*: Gauteng core, 2018 and 2022 cohorts.

Figure 19 shows the structure of the generated feeder network: 200 feeders, each sending students to a median of 11 Grade 8 destinations at 50 students per feeder. The matrix is sparse, which is the source of the finite-sample noise documented in Figure 18 at $\psi = 0$. The bounding results in that figure are therefore conditioned on a network that genuinely replicates the empirical sparsity the estimator faces.

Figure 20 shows the endogenous-mobility channel at $\psi = 1$: $\text{corr}(z_i, q_k) = 0.658$, a positive slope confirming the DGP embeds the threat. The flat reference at $\psi = 0$ yields $\text{corr}(z_i, q_k) = 0.001$. The empirical data correspond to $\rho_{\text{emp}} \approx 0.021$ (dashed horizontal in the figure), which lies far below the $\psi = 1$ scenario. The bounding sweep’s “does not bind” verdict in Figure 18 is therefore not an artefact of a weak or absent threat: the channel is calibratable in the DGP, and the empirical data sit in a region where it does not bind.

Limited-mobility DGP: each feeder reaches few destinations

Median 11 destinations per feeder; network density 21.8% — sparse bipartite structure visible

Feeders: 200 | Destinations: 50

Median destinations per feeder: 11

Network density: 21.8%

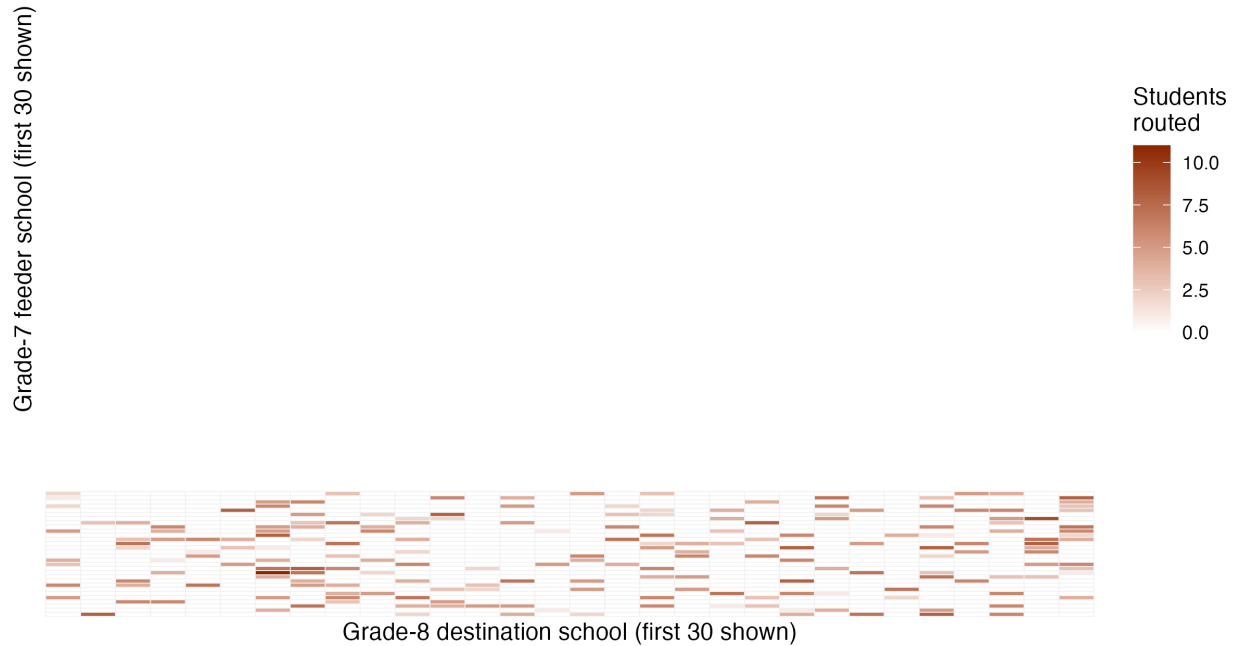


Figure 19: DGP face-validity: limited mobility in the generated feeder-to-destination network. Each cell in the bipartite matrix records the number of students flowing from a Grade 7 feeder school (rows) to a Grade 8 destination school (columns). The network is sparse by construction: 200 feeder schools, median 11 destinations per feeder, 50 students per feeder. Most cells are empty; each feeder reaches only a narrow slice of destinations. The bounding sweep in [Appendix L.1](#) uses this same observed network density as its baseline scenario. *Sample*: generated data from the Monte Carlo design of §5.1 of the main text, not observed students.

Endogenous-mobility DGP: channel active at $\psi = 1$, flat at $\psi = 0$

Realised $\text{corr}(z_i, q_k) = 0.658$ at $\psi = 1$; 0.001 at $\psi = 0$

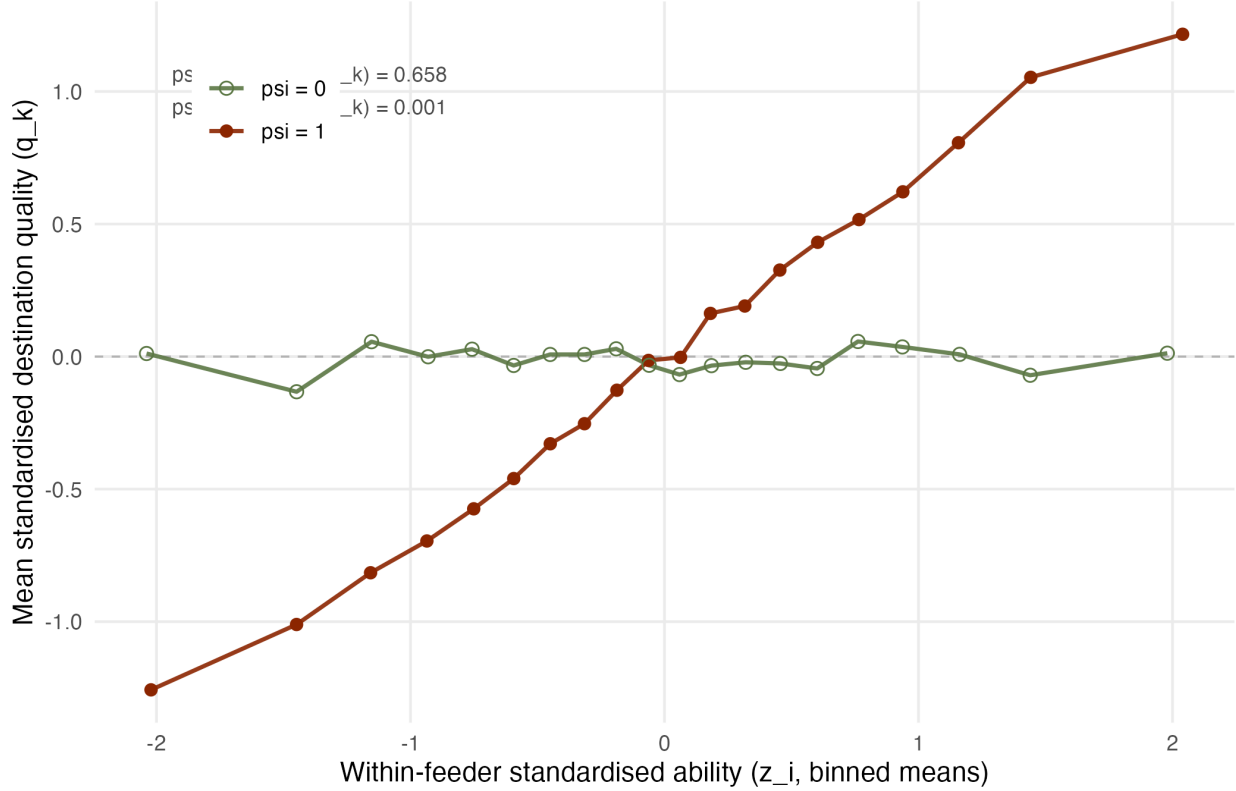


Figure 20: DGP face-validity: endogenous mobility at $\psi = 1$ vs $\psi = 0$. Binned means of standardised destination school quality q_k (the standardised true school fixed effect γ_k) against standardised within-feeder ability z_i . At $\psi = 1$ (orange), the slope is positive: $\text{corr}(z_i, q_k) = 0.658$, confirming that the channel is active and correctly implemented. At $\psi = 0$ (grey), the slope is flat: $\text{corr}(z_i, q_k) = 0.001$. The empirically observed within-feeder proxy $\rho_{\text{emp}} \approx 0.021$ (dashed line) sits far below the $\psi = 1$ scenario; the bounding sweep in [Appendix L.1](#) shows the empirical data correspond to ψ well below the 5% bias threshold. This figure plots $\psi = 1$ to demonstrate the mechanism is implemented correctly and is calibratable; it does not describe the empirical setting. *Sample:* generated data from the Monte Carlo design of §5.1 of the main text, not observed students.

L.2: Heaping 2D bounding (Gr7 $\times$ Gr8 surface)

The headline heaping MC of §5.1 of the main text applies threshold-pushing to the Grade 7 mark only and uses an IV stipulated to correlate 0.85 with the *unheaped* truth (the IV here is a DGP comparator within the synthetic experiment; the empirical attenuation correction uses structural de-heaping, §5.3 of the main text). Either choice could in principle do the rank-recovery work mechanically. To bound that concern, we sweep a 2D grid of target heaping indices on the (Gr7 baseline, Gr8 endline) marks, with both axes spanning $\{1.0, 1.5, 2.0, 2.4, 3.0, 4.0, 5.0\}$ (200 replications per cell, same DGP and estimators as §5.1 of the main text) and measure how OLS $\hat{\beta}_1$ moves across the 49-cell surface relative to the *infeasible oracle* benchmark estimated on the unheaped data. The upper extreme of the realised sweep is 2.44 at the 40-mark threshold, 9 per cent above the empirical Gr7-2022 heaping of 2.24, set deliberately high so a bias breakpoint can be located if one exists. The pure heaping-induced shift $(\hat{\beta}_1^{\text{OLS}} - \hat{\beta}_1^{\text{oracle}})/\beta_1$ ranges from -0.24% to $+0.71\%$ across the entire surface; the worst-case bias magnitude is 0.71%, well short of the 1% threshold, let alone the 5% that would matter for any downstream interpretation. The 5% bias contour does not exist within the testable range. The two heaping channels operate in opposite directions: Gr7 (RHS) heaping shrinks the within-school variance of the regressor more than its covariance with Grade 8 and biases $\hat{\beta}_1 = \widehat{\text{Cov}}/\widehat{\text{Var}}$ *up*, with worst case $+0.71\%$ at the corner where Gr7 is maximally heaped and Gr8 is clean; Gr8 (LHS) heaping creates non-classical conditional-mean mass-shifting and biases $\hat{\beta}_1$ *down*, with worst case -0.24% at the mirror corner where only Gr8 is heaped. Symmetric application of the mechanism to both marks therefore produces a smaller net bias because the two channels partially offset. At the empirically observed (Gr7, Gr8) heaping indices, (1.93, 1.48) in 2018 and (2.24, 1.81) in 2022, the implied OLS bias is $+0.40\%$ in 2018 and $+0.55\%$ in 2022 (Figure 21), smaller than the variance-reduction and specification effects audited in §4.2 of the main text.²

L.3: Peer-effects variants and bounding

The sorting MC of §5.1 of the main text uses a linear-in-means peer-effect specification with own-Gr8-class peer ability. Three functional-form variants test channels that bypass the $1/n$ structural attenuation: heterogeneous λ_j across feeder schools (the $j \times k$ channel), multiplicative own-ability $\times$ peer-ability, and a top- k share-above-cutoff threshold. Across the (ρ, λ) grid that produces the sorting Monte Carlo summary (Table 6 of the main text), all three variants track the linear-in-means baseline cell-for-cell on rank correlation, RMSE, variance ratio and coverage; ρ remains the only dimension that moves the metrics. A bounding sweep at $\rho = 0.6$ extends the λ grid to $\lambda \in [0, 4]$ and identifies, for each variant, the critical λ^* at which $\text{RMSE}(\hat{\alpha}_j)$ inflates by more than 5% relative to the no-peer baseline: $\lambda^* \approx 0.18$ (heterogeneous λ_j), $\lambda^* \approx 0.19$ (multiplicative own-ability $\times$ peer-ability), and $\lambda^* > 4$ for the top- k threshold, which reaches only 3.7% RMSE inflation at $\lambda = 4$. The linear-in-means peer-effect literature reports estimates in $\lambda \in [0.05, 0.20]$, so the empirically plausible range sits at or below the 5% bias threshold across every variant. An empirical companion adds a Grade 8 schoolmate leave-out baseline-mark control to the headline TWFE on the Gauteng-core sample: the school fixed effects keep a Pearson correlation with the baseline specification of at least 0.9999 in all four (subject, cohort) cells; the largest mean absolute level shift in any cell is 0.0021 z-units and the largest single shift 0.13 z-units. Class-level peer-effect channels remain tested only in the Monte Carlo because the source data carry no Grade 8 class identifier.

Figure 22 confirms that positive assortative sorting is present in the generated data. At $\rho = 0.6$, the leave-out peer-ability mean rises with own ability (realised corr = 0.136, in units of within-school standardised ability); at $\rho = 0$ the slope is flat (corr = 0.027). The bounding sweep in Appendix L.3 uses $\rho = 0.6$ as the headline stress scenario, so it tests a non-trivial degree of sorting, not a synthetic scenario where the threat is absent.

L.4: Attrition K-anchored bounding

The attrition MC of §5.1 of the main text reports the seven-scenario summary at fixed empirical anchor severity. The empirical anchor itself is small: the observed Pearson correlation between school linkage rates

²Two single-axis adversarial variants were also run: a symmetric application to both marks, and a heaping-contaminated instrument whose first-stage informativeness erodes only from 0.850 to 0.849 across the grid.

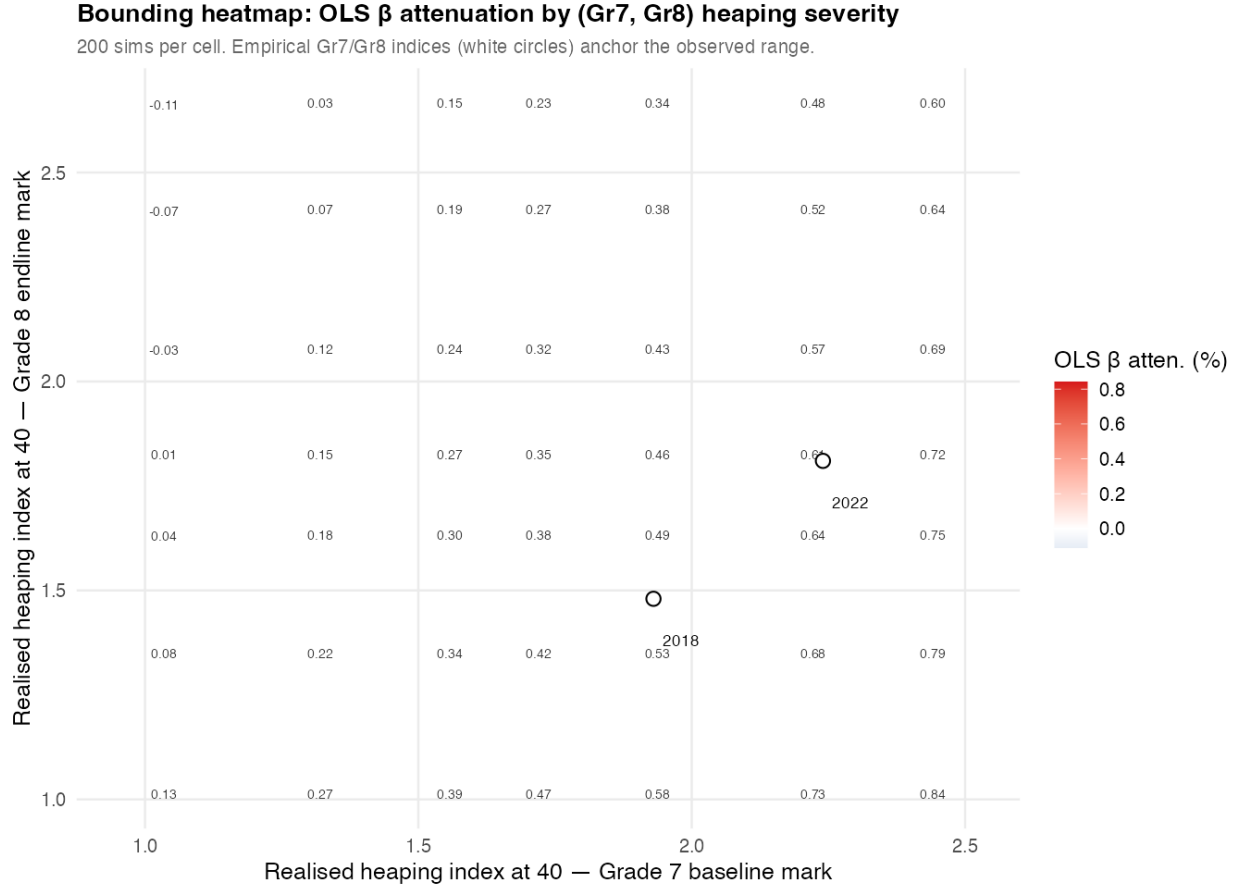


Figure 21: Bounding heatmap. Pure heaping-induced shift $(\hat{\beta}_1^{\text{OLS}} - \hat{\beta}_1^{\text{oracle}})/\beta_1$ (as a percentage of the true persistence parameter) across a 49-cell sweep of target heaping indices on the (Gr7 baseline, Gr8 endline) marks; axes show the mean realised heaping index at the 40-mark threshold. White circles mark the empirically observed (Gr7, Gr8) heaping indices for the 2018 and 2022 cohorts. The shift ranges from -0.24% to $+0.71\%$ across the entire surface; the worst-case bias magnitude is 0.71% , at the heaviest-Gr7, cleanest-Gr8 corner. The 1% and 5% bias contours do not exist within the testable range. 200 replications per cell. *Sample*: simulated data calibrated to the Gauteng core, 2018 and 2022 cohorts.

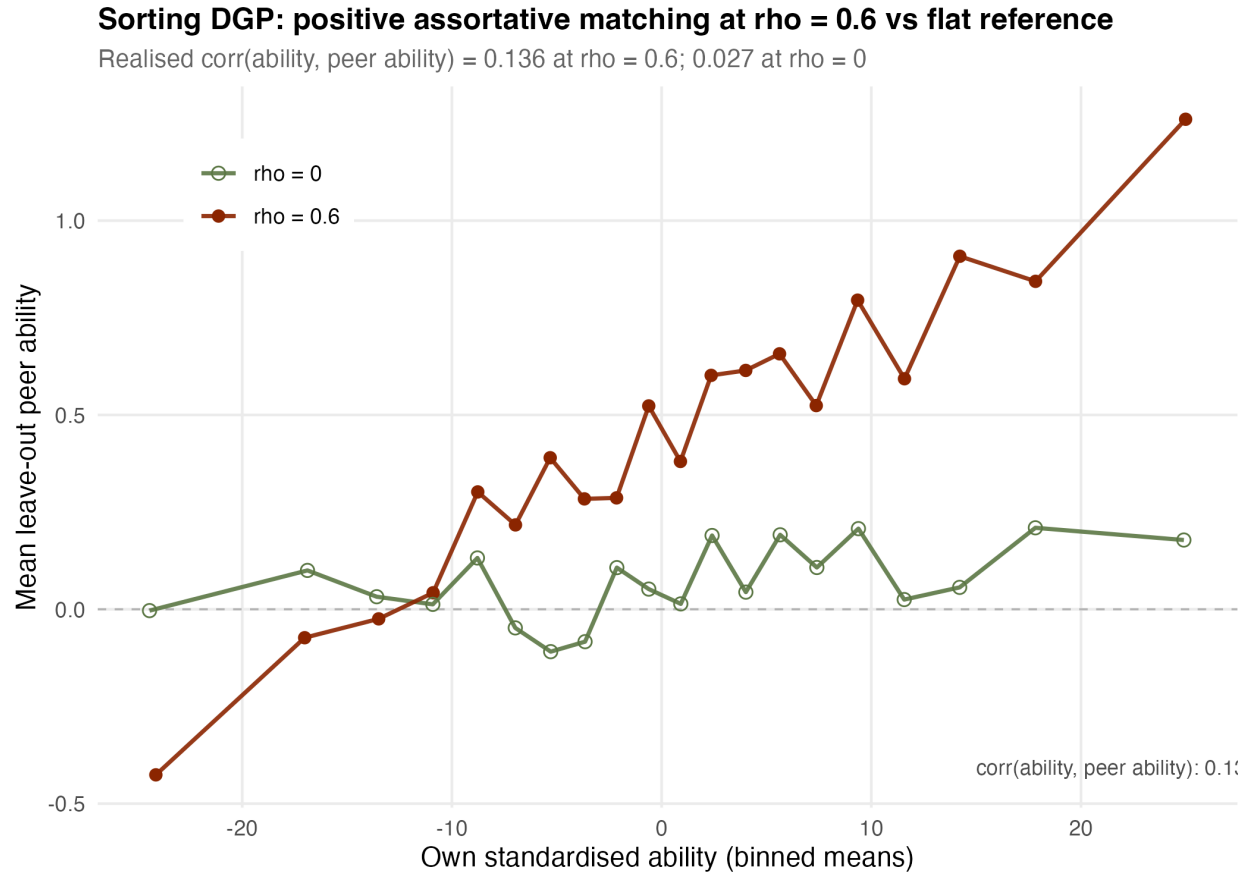


Figure 22: DGP face-validity: positive assortative sorting in generated peer groups. Binned means of leave-out peer ability against own standardised ability, at sorting intensity $\rho = 0.6$ (orange) and the flat reference $\rho = 0$ (grey). The realised within-class $\text{corr}(\text{own ability}, \text{peer ability})$ is 0.136 at $\rho = 0.6$ and 0.027 at $\rho = 0$. The positive assortative slope at $\rho = 0.6$ confirms that peer-group composition is informative about own ability in the generated data; the flat reference confirms the channel switches off cleanly at $\rho = 0$. The bounding results in [Appendix L.3](#) are therefore conditioned on a live sorting threat. *Sample:* generated data from the Monte Carlo design of §5.1 of the main text, not observed students.

and estimated school effects is 0.021 (Table 5 in Appendix A), pooled over the twelve province-cohort fits; within fits it averages -0.046 . To bound the heterogeneous-treatment-effect $\times$ selective-attribution channel that Appendix A.6 flags as the substantive remaining concern, we run a K-anchored sweep at $1\times, 2\times, 5\times, 10\times$ empirical severity, crossed with a treatment-effect heterogeneity grid $\delta_{\text{het}} \in \{0, 0.25, 0.5\}$. At the empirical anchor ($K = 1\times$), the maximum Q5–Q1 bias spread is 0.052 marks; at $K = 10\times$ severity with $\delta_{\text{het}} = 0.5$, the maximum spread is 0.118 marks (well under 2% of $\sigma_\tau \approx 6$). The 5%- σ_τ contour (≈ 0.3 marks) is not crossed anywhere on the $K \times \delta_{\text{het}}$ grid. Pushing the across-school differential above that contour therefore requires a selection severity more than ten times the pooled correlation, and more than four times the within-fit one.

Attrition DGP: non-random linkage gradient present in generated data

Observed linkage rate 80.0%; $\text{corr}(\text{linked}, \text{ability}) = 0.094$, $\text{corr}(\text{linked}, \text{alpha}) = 0.008$

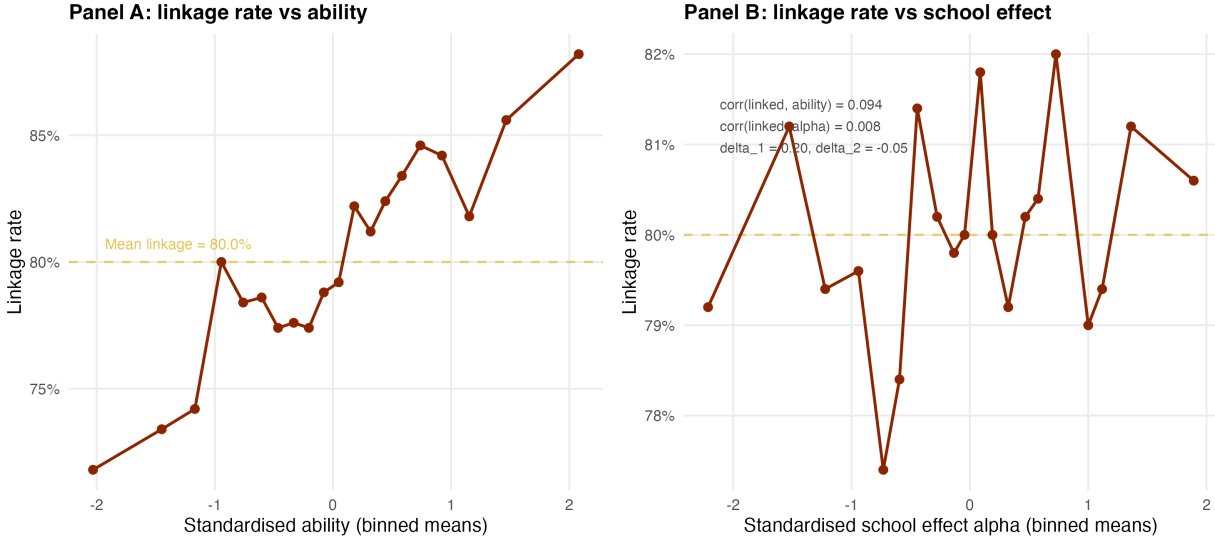


Figure 23: DGP face-validity: selective attrition in the generated linkage process. Left panel: empirical linkage rate (binned means) against standardised own ability; right panel: linkage rate against standardised school effect α . Both gradients are non-flat, confirming that the DGP embeds the attrition threat at the headline severity parameters. The observed Pearson correlation between generated school linkage rates and school effects is 0.008 (standardised units); the empirical anchor in the live data is 0.021 (Table 5), pooled over the twelve fits; within fits it averages -0.046 . The overall linkage rate in the generated draw is 80%, matching the empirical level. *Sample*: generated data from the Monte Carlo design of §5.1 of the main text, not observed students.

Figure 23 confirms that the attrition DGP embeds a non-flat selection gradient against both ability (left panel) and school quality α (right panel). The generated $\text{corr}(\text{linkage rate}, \text{school effect})$ is 0.008 in standardised units; the empirical anchor from Table 5 is 0.021, pooled over the twelve fits. The K-anchored sweep in Appendix L.4 runs at multiples of that empirical severity, so the “does not bind” verdict in the sweep is evaluated against a threat that is present in the DGP and anchored to the data.

L.5: Heaping–quality correlation bounding

The heaping MC of §5.1 of the main text and the L.2 severity sweep heap every school with the same push probability, so per-school severity is independent of the true school effect. The data are not like that: heaping is heaviest at the lowest-scoring schools. Across schools the per-school heaping share correlates with the raw mean Grade-7 mark at Spearman -0.62 in 2018 and -0.67 in 2022; against the recovered school effect the correlation is smaller and of the opposite sign, 0.27 and 0.40 , because the recovered effect and the raw mark are themselves negatively correlated. A school low on the raw mark but high on value-added heaps most. Since heaping only raises marks, severity concentrated on the weakest schools could differentially inflate the low end of the school-effect distribution and reorder schools in a way uniform severity cannot.

To bound that channel, we make each school’s push probability a function of school quality, on two axes run separately: decreasing in the raw mean Grade-7 mark, and increasing in the true school effect. A mean-preserving multiplier holds average severity fixed at the headline level while a single parameter steepens the severity–quality gradient, so the sweep isolates the correlation from the severity that [Appendix L.2](#) already bounds. The gradient is calibrated to the data and pushed past it, to a bottom-to-top quintile ratio of 9 against the 5.6 the marks show. Across the sweep the rank correlation with the truth stays within 0.0004 of the clean-mark oracle, the persistence slope moves by under 0.4%, and the estimated school-effect spread never compresses (its ratio to the truth holds at 1.02 or above); the bias in the bottom quality quintile, where the inflation would concentrate, stays under 0.10 of a mark against $\sigma_\tau \approx 6$. Directional heaping tied to school quality, at the gradient the marks show and steeper, does not reorder schools (300 replications per cell).

Appendix M: Revealed-preference test, extended results

Scope restriction and sample construction

The revealed-preference exercise runs on the Gauteng core, the mathematics sample used from §3 of the main text through §5 of the main text, where the Grade 7 to Grade 8 feeder network is densest and the two-way clustering has the most origin and destination schools to lean on. English first additional language is excluded because within-year reliability falls below the bar set in §4.2 of the main text. Estimating the same test province by province across all six provinces places the pooled coefficient below the Gauteng-core value (Table 34), so §6.1 of the main text reports the Gauteng core throughout. Table 35 traces the sample from all Grade 7 to Grade 8 mathematics movers to this headline test: the join to the calibrated estimates and the mathematics restriction, a destination-cohort floor of at least five mathematics peers, complete demographic controls, and the final restriction to the Gauteng core, which leaves 221,093 student-subject-cohort observations.

Table 34 reports the achievement-margin coefficient in each province. The calibrated signal is strongest in Gauteng, holds in Mpumalanga, and thins to near zero in Eastern Cape and KwaZulu-Natal; pooling all six provinces roughly halves the coefficient relative to Gauteng. The Gauteng-core headline is therefore the sample where the sorting signal is cleanest to read, not a province selected on the strength of its result.

Table 34: Revealed-preference coefficient by province. The wedge $\hat{\alpha}_j$ as a predictor of the destination cohort’s Grade 8 peer mean, estimated separately in each of the six provinces and pooled across all six, under leave-one-out (LOO) and leave-own-feeder-out (LOFO) peer constructions. Coefficients in within-cohort standard-deviation units. The final column is the LOFO coefficient on the destination’s raw Grade 7 intake, the construction diagnostic of §6.1.4 of the main text.

Province	n	Peer G8 (LOO)	Peer G8 (LOFO)	t (LOFO)	Raw G7 (LOFO)
EC	162,355	+0.068	+0.013	0.77	+0.064
GT	220,973	+0.166	+0.135	7.22	+0.077
LP	182,360	+0.107	+0.043	2.55	−0.055
MP	106,955	+0.144	+0.091	4.29	+0.078
NW	91,119	+0.117	+0.047	2.38	−0.002
ZN	254,501	+0.068	+0.013	0.83	+0.013
Six-province pooled	1,018,263	+0.117	+0.064	8.38	+0.030

Sample: Mathematics, 2018 and 2022 cohorts. Each province is estimated from the six-province pool, standardised over the pooled cohort-by-subject standard deviation, so the Gauteng cell sits a few thousandths below the own-sample Gauteng-core headline of §6.1 of the main text; the final row pools all six provinces.

Table 35: Sample attrition: mathematics, Gauteng core. *Sample*: each step’s own, from all switchers (all provinces and subjects) to the headline sample: Mathematics; Gauteng core; 2018 and 2022 cohorts.

Step	n
All switchers (non-missing identifiers, Grade 7 mark; all provinces and subjects)	1,943,200
After alpha join — restricts to GT/MP/NW, Mathematics only	432,630
Destination cohort at least 5 mathematics peers	432,630
Listwise-complete on demographic controls	423,149
Restrict to Gauteng core (headline sample)	221,093

Table 36: Pairwise correlations among the three origin-school signals (school-level, mathematics, Gauteng core). *Sample*: origin schools, at school level; Mathematics; Gauteng core.

Pair	Correlation
$\hat{\alpha}_j$ vs $\bar{y}_j^{G7}$	-0.469
$\hat{\alpha}_j$ vs $\bar{y}_j^{G8}$	0.595
$\bar{y}_j^{G7}$ vs $\bar{y}_j^{G8}$	0.232

The pattern of predictor correlations matches the school-level correlations reported in Table 36: $\hat{\alpha}_j$ is negatively correlated with the raw input signal and positively correlated with the raw output signal, reflecting the difficulty interpretation the calibration is designed to capture.

Horse-race specification

Figure 24 assembles all six single-predictor margins in one view, the Grade 8 achievement block (rows a to c) above the Grade 7 intake block (rows d to f), placing the achievement result of §6.1.3 of the main text beside the intake result of §6.1.4 of the main text. The intake block reads the construction artefact at a glance: the wedge sits at -0.069 within-cohort SD (row d), the own-feeder artefact taken apart in §6.1.4 of the main text, while the raw Grade 7 mark sits at $+0.310$ (row e), the mark-on-mark matching the wedge nets out, and the raw Grade 8 mark sits at $+0.223$ (row f), a weak residual far below its own-margin Grade 8 value. The joint specification the figure also plots is a bad-control exercise taken up in the robustness checks of §6.1.5 of the main text, not a horse race: conditioning a measure on its own inputs flips its signs.

Entering all three signals simultaneously in a single regression partially orthogonalises each coefficient against the other two. By the bridge identity of §3 of the main text, $\bar{y}_j^{G8}$ contains $\hat{\alpha}_j + \beta_1 \bar{y}_j^{G7}$ up to estimation noise, so conditioning on it mechanically reverses the calibrated partial coefficients: with $\hat{\alpha}_j$ negatively correlated with $\bar{y}_j^{G7}$ and positively correlated with $\bar{y}_j^{G8}$ (Table 36), both calibrated cells in Table 37 flip sign relative to their single-predictor counterparts, the $\hat{\alpha}_j$ coefficient turning negative on the Grade 8 outcome and positive on the Grade 7 outcome. The single-predictor specifications remain the interpretable cells, and the construction-robust decomposition is in Appendix M.

Table 37: Horse-race specification: all three origin-school signals entered simultaneously. Same controls and clustering as the revealed-preference headline table in the main text. * $p < 0.001$. ***Sample*: the headline sample of Table 9 of the main text: Mathematics; Gauteng core; 2018 and 2022 cohorts.

Origin-school signal $\mathcal{S}_j$	Dest. peer G7 mark (z)	Dest. peer G8 mark (z)
$\bar{y}_j^{G7}$ (mean Grade 7 mark, z)	+0.445*** (0.036)	−0.252*** (0.021)
$\bar{y}_j^{G8}$ (mean Grade 8 mark, z)	+0.056 (0.034)	+0.783*** (0.027)

(continued)

Origin-school signal $\mathcal{S}_j$	Dest. peer G7 mark (z)	Dest. peer G8 mark (z)
$\hat{\alpha}_j$ (wedge, z)	+0.234*** (0.044)	−0.326*** (0.029)

Dropping the raw Grade 8 mean, the one signal that contains the wedge by construction, leaves a fit that holds the raw Grade 7 mean and the wedge together (Table 38). The two correlate at -0.469 across schools, far from the near-collinearity of the three-signal fit. On the Grade 8 leave-own-feeder-out peer mean, with the student’s own mark and the controls, both enter positively: the raw Grade 7 mean at $+0.181$ ($t = 8.1$) and the wedge at $+0.269$ ($t = 9.9$). Sorting on the calibrated learning level fixes the ratio of the two coefficients in advance, since the level weights the raw mean by the calibration slope: 0.82 in 2018 and 0.64 in 2022. The estimated ratios, 0.75 and 0.62 , are within sampling error of those values ($p = 0.23$ and $p = 0.70$). Each signal enters with the weight the calibration gives it: students sort on the calibrated scale.

Table 38: The raw Grade 7 mean and the wedge enter in the ratio the calibration assigns

	Leave-own-feeder-out			Own-pupil-out
	2018 (1)	2022 (2)	Pooled (3)	Pooled (4)
Raw Grade 7 mean	0.216*** (0.033)	0.171*** (0.025)	0.181*** (0.022)	0.217*** (0.021)
Wedge $\hat{\alpha}_j$	0.288*** (0.037)	0.277*** (0.034)	0.269*** (0.027)	0.324*** (0.025)
Own Grade 7 mark	0.068*** (0.008)	0.066*** (0.010)	0.069*** (0.009)	0.068*** (0.008)
Coefficient ratio	0.75	0.62	—	—
Ratio the calibration implies	0.82	0.64	—	—
p , equal ratios	0.23	0.70	—	—
Student and school controls	Yes	Yes	Yes	Yes
Province-by-cohort effects	Yes	Yes	Yes	Yes
Observations	94,972	126,001	220,973	221,093

Notes: Columns (1) to (3) regress the leave-own-feeder-out Grade 8 peer mean, and column (4) the own-pupil-out Grade 8 peer mean, on the origin school’s raw Grade 7 mean and wedge, with the student’s own Grade 7 mark, student and school controls (sex, home language, population group, over-age, poverty quintile, urban or rural location, no-fee status) and province-by-cohort effects. Coefficients are in within-cohort SD of the outcome per SD of the signal; standard errors, in parentheses, are clustered by primary and by secondary school. The coefficient ratio divides the raw-mean coefficient by the wedge coefficient; the ratio the calibration implies is the weight the calibrated learning level gives the raw Grade 7 mean relative to the wedge in that cohort, and the p -value is a Wald test that the two are equal. The pooled columns mix two cohorts with different implied ratios, so they carry no ratio. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an em dash. Sample: Mathematics, Gauteng core, 2018 and 2022 cohorts: the leave-own-feeder-out sample of Table 42, 220,973 student-subject-cohort observations (94,972 in 2018, 126,001 in 2022); in column (4), the headline sample, 221,093.

Bootstrap confidence intervals

First-stage uncertainty in $\hat{\alpha}_j$ arises because the calibrated learning levels are estimated, not observed. To propagate this uncertainty into the second-stage sorting regression, we implement a paired-cluster bootstrap with $B = 200$ replicates. In each replicate, origin-school clusters are resampled with replacement and a perturbed $\hat{\alpha}_b \sim \mathcal{N}(\hat{\alpha}, \widehat{\text{se}}_{\alpha})$ is drawn independently per origin school, re-standardised within cohort on the resample, and substituted as the regressor in the second-stage equation estimated on the resampled clusters. The 95% confidence intervals reported in §6.1 of the main text are the 2.5th and 97.5th percentiles of the resulting coefficient distribution across replicates.

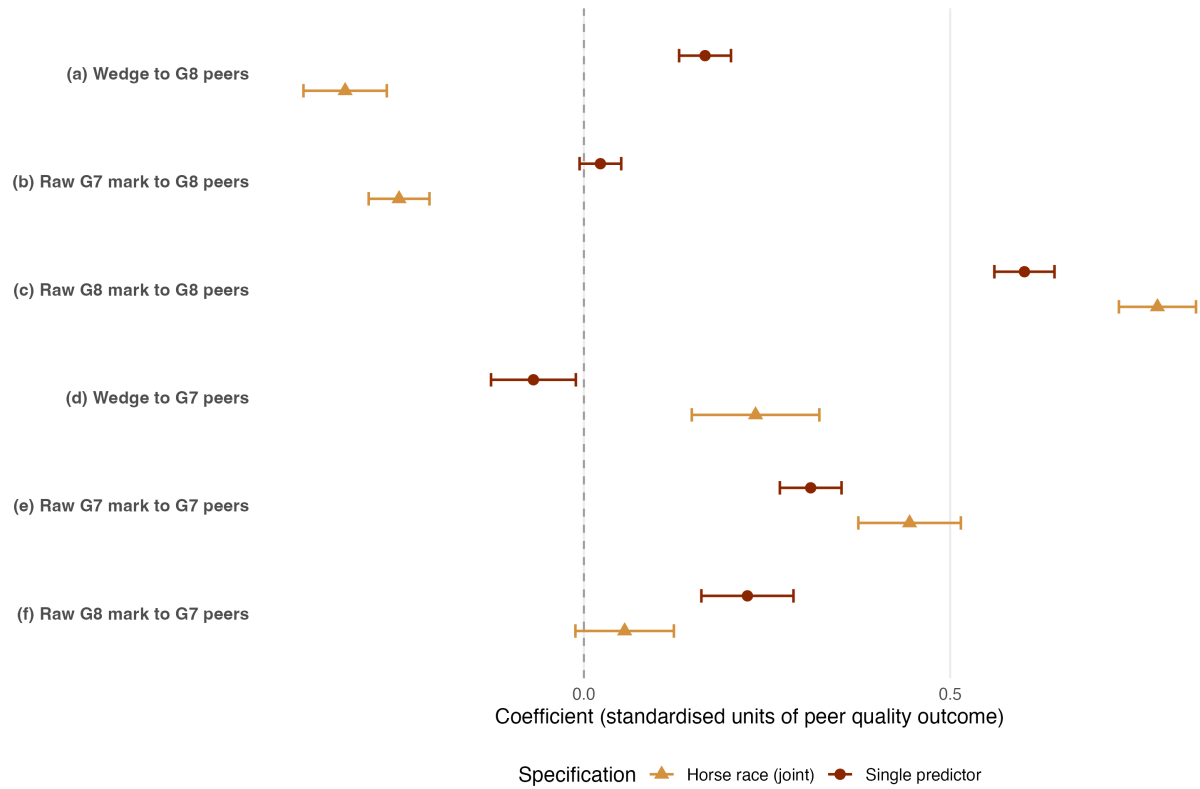


Figure 24: **Each origin signal as a predictor of the destination cohort’s Grade 7 and Grade 8 peer means.** Coefficient and 95% confidence interval (from two-way-clustered standard errors) on each of the three origin-school signals, under single-predictor fits (rows a to f) and a joint specification entering all three signals together. Rows a to c use the destination Grade 8 achievement margin; rows d to f use the destination Grade 7 intake margin. The single-predictor rows are the interpretable ones. The joint specification is a bad-control exercise, not a horse race: because the wedge is by construction the Grade 8 mark net of a fraction of the Grade 7 mark, conditioning on both raw means reverses its coefficients’ signs. Table 37 reports the joint coefficients in full. **Sample:** 221,093 student-subject-cohort observations; 1,371 origin / 883 destination schools; Gauteng core; Mathematics; 2018 and 2022 cohorts.

Table 39: Bootstrap confidence intervals propagating first-stage $\hat{\alpha}$ estimation uncertainty. Point estimate equals the second-stage coefficient using the empirical $\hat{\alpha}$. Bootstrap SE and 95% CI from $B = 200$ paired-cluster replicates. *Sample:* the headline sample of Table 9 of the main text: Mathematics; Gauteng core; 2018 and 2022 cohorts.

Outcome	Point est.	Boot. SE	95% CI
Dest. peer G7 mark (z)	-0.069	0.020	[-0.107, -0.032]
Dest. peer G8 mark (z)	0.165	0.013	[0.136, 0.189]

Cross-cohort alpha robustness

The pooled cross-cohort estimate in §6.1 of the main text combines two directional sub-analyses: (i) 2018 choices predicted by 2022 $\hat{\alpha}$ ($n = 206,628$ restricted to 2018 cohort movers; destination Grade 8 estimate $+0.160$, $t = 7.5$), and (ii) 2022 choices predicted by 2018 $\hat{\alpha}$ (destination Grade 8 estimate $+0.078$, $t = 3.2$). Both directional estimates are positive and statistically distinguishable from zero on the destination Grade 8 peer-mark outcome. The pooled estimate of $+0.108$ ($t = 5.5$) is a precision-weighted combination; because the 2018-choices/2022-alpha cell has a larger within-province mover sample, it receives higher weight. The destination Grade 7 peer-intake outcome is estimated near zero in both directions, consistent with the interpretation that cross-cohort noise attenuates the input-selectivity channel more than the wedge channel.

Table 40: Cross-cohort alpha robustness. Each row substitutes the opposite cohort's $\hat{\alpha}$ to sever within-cohort mechanical correlation. SE clustered two-way at origin Grade 7 and destination Grade 8 school. *Sample:* movers in the headline sample whose origin school has a wedge in the other cohort; Mathematics; Gauteng core; n by row in the last column.

Specification	Outcome	$\hat{\beta}$	SE	t	n
cross-cohort (pooled)	Dest. peer G7 (z)	0.027	0.026	1.06	206,628
cross-cohort (pooled)	Dest. peer G8 (z)	0.108	0.020	5.47	206,628
cross-cohort (2018 choices, 2022 alpha)	Dest. peer G7 (z)	0.071	0.029	2.45	91,832
cross-cohort (2018 choices, 2022 alpha)	Dest. peer G8 (z)	0.160	0.021	7.48	91,832
cross-cohort (2022 choices, 2018 alpha)	Dest. peer G7 (z)	-0.011	0.028	-0.41	114,796
cross-cohort (2022 choices, 2018 alpha)	Dest. peer G8 (z)	0.078	0.024	3.22	114,796

Mechanical structure of the G7 sign reversal

The negative bottom-left estimate of the revealed-preference headline table in the main text (the wedge's coefficient on the destination Grade 7 peer-intake outcome) is a corollary of the school-level joint distribution of the three origin signals, not an independent empirical finding. The wedge $\hat{\alpha}_j$ is, at the school level, negatively correlated with the raw Grade 7 mean ($r = -0.469$ between $\hat{\alpha}_j$ and $\bar{y}_j^{G7}$ on the restricted-scope sample, consistent with the wider §3.4 of the main text estimate). Origins with higher $\hat{\alpha}_j$ are predominantly schools with harder assessments whose raw Grade 7 marks are depressed by difficulty rather than by weaker latent learning. Any destination-side attribute that tracks the raw Grade 7 input level positively will, by that correlation, track $\hat{\alpha}_j$ with the opposite sign. The destination-side projection of the origin-side $\hat{\alpha}_j/\bar{y}_j^{G7}$ correlation is therefore mechanical.

The substantive independent content of the revealed-preference headline table in the main text lies elsewhere: in the *positive* bottom-right estimate (the wedge's coefficient on the destination Grade 8 peer-mark outcome) and in the *opposition* of the wedge's two estimates across destination outcomes. Neither is fixed by the school-level $\hat{\alpha}_j/\bar{y}_j^{G7}$ correlation. The cross-cohort design of [Appendix M](#) is the direct test: substituting the opposite

cohort's $\hat{\alpha}$ weakens the mechanical correlation between estimation stages and, exactly as the mechanical reading predicts, attenuates the destination Grade 7 estimate to near zero while leaving the destination Grade 8 estimate intact.

Same-feeder exposure and the leave-own-feeder-out re-estimate

The destination peer means in the revealed-preference headline table in the main text are leave-one-out at the *student* level: a mover's own mark is excluded, but her primary-school feeder-mates who moved to the same destination are not. Because $\hat{\alpha}_j$ is estimated on the same cohort's Grade 8 marks, same-feeder co-movers carry origin $\hat{\alpha}_j$ into the outcome mechanically. Table 41 reports the exposure: the share s_i of each student's leave-one-out peer pool drawn from her own feeder.

Table 41: Same-feeder share s_i of the leave-one-out destination peer pool. Sample: $n = 221,093$ student-subject-cohort observations (the headline estimation sample); mathematics; Gauteng core.

Scope	Mean	Median	p90	Share $s_i = 0$	Share $s_i > 0.5$
Gauteng core	0.150	0.108	0.359	0.097	0.038
2018 cohort	0.157	0.112	0.372	0.104	0.048
2022 cohort	0.145	0.105	0.349	0.091	0.031

Table 42 reports the leave-own-feeder-out (LOFO) re-estimates, excluding all same-feeder students from the peer mean and mirroring the headline's five-peer destination floor (sensitivity at ≥ 1 and ≥ 10 cross-feeder peers shown). The exact decomposition rows scale the same-feeder and cross-feeder components of the leave-one-out outcome by the *same* within-cohort standard deviation, so the two $\hat{\alpha}_j$ coefficients sum to the headline coefficient by OLS linearity. The γ_k -netted rows remove the destination effect, the first stage's destination fixed effects demeaned within province $\times$ cohort, which carries a destination's assessment difficulty together with its teaching (§6.1.5 of the main text). The cross-cohort $\times$ LOFO row composes this exclusion with Appendix M's design, severing the estimation-overlap and composition channels simultaneously. The paired-cluster bootstrap for the LOFO Grade 8 cell ($B = 999$; origin clusters resampled with replacement, $\hat{\alpha}$ redrawn per replicate) gives the 95% interval reported below.

Table 42: Leave-own-feeder-out re-estimates and the exact same-/cross-feeder decomposition. Same controls and clustering as the revealed-preference headline table in the main text. Sample: $n = 220,973$ (destinations with ≥ 5 cross-feeder mover peers on both margins; 99.9% of the headline sample); mathematics; Gauteng core; two-way clustered (origin, destination). *** $p < 0.001$.

Cell	Estimate (SE)	t	n
$\hat{\alpha}_j \rightarrow \text{dest. G8, leave-one-student-out (headline)}$	+0.165*** (0.018)	+9.2	221,093
$\hat{\alpha}_j \rightarrow \text{dest. G8, LOO on LOFO subsample}$	+0.165*** (0.018)	+9.1	220,973
$\hat{\alpha}_j \rightarrow \text{dest. G8, leave-own-feeder-out}$	+0.136*** (0.018)	+7.4	220,973
$\hat{\alpha}_j \rightarrow \text{dest. G8, } \gamma_k\text{-netted LOO}$	+0.273*** (0.020)	+13.8	221,093
$\hat{\alpha}_j \rightarrow \text{dest. G8, } \gamma_k\text{-netted LOFO}$	+0.231*** (0.020)	+11.3	220,973
$\hat{\alpha}_j \rightarrow \text{dest. G7, leave-one-student-out (headline)}$	-0.069 (0.030)	-2.3	221,093
$\hat{\alpha}_j \rightarrow \text{dest. G7, leave-own-feeder-out}$	+0.102*** (0.029)	+3.6	220,973
$\bar{y}_j^{G7} \rightarrow \text{dest. G8, LOFO (placebo)}$	+0.020 (0.015)	+1.4	220,973
$\bar{y}_j^{G7} \rightarrow \text{dest. G7, LOFO (like-with-like)}$	+0.109*** (0.022)	+4.9	220,973
Decomposition: same-feeder component (same-SD scaling)	+0.052*** (0.013)	+4.1	221,093
Decomposition: cross-feeder component (same-SD scaling)	+0.113*** (0.019)	+5.9	221,093
Sensitivity: ≥ 1 cross-feeder peers	+0.136*** (0.018)	+7.4	221,071
Sensitivity: ≥ 10 cross-feeder peers	+0.136*** (0.019)	+7.3	220,129

(continued)

Cell	Estimate (SE)	t	n
Cross-cohort $\times$ LOFO, $\hat{\alpha}_j \rightarrow$ dest. G8 (pooled)	+0.095*** (0.020)	+4.8	206,545

Dropping every same-feeder peer leaves only the achievement result standing, exactly as the bridge predicts. On the leave-own-feeder-out peer means, recomputed after dropping each mover’s own feeder-mates and retaining destinations with at least five cross-feeder peers (99.9% of the sample, $n = 220,973$), the wedge still predicts Grade 8 achievement at +0.136 ($t = +7.4$; bootstrap 95% interval [+0.107, +0.161]), while the negative intake coefficient flips to +0.102 ($t = +3.6$). The achievement result survives the one channel through which the construction could have manufactured it; the intake result does not.

A raw-mark benchmark points the same way without assuming the result. Because flows are feeder-clustered, an allocation on raw Grade 7 marks alone, with no household sorting, would still produce a positive same-feeder cell, while its cross-feeder prediction is zero only insofar as school-level raw marks carry no cross-school achievement signal. Both halves hold in the data: cross-feeder matching on raw marks stays real (raw Grade 7 on cross-feeder Grade 7 intake, +0.109, $t = +4.9$) yet predicts nothing about achievement (raw Grade 7 on cross-feeder Grade 8, +0.020, $t = +1.4$), where the wedge remains positive. The same-feeder share of the headline coefficient, 32%, is therefore an upper bound on its mechanical fraction, since it also holds any genuine same-feeder behaviour; the full decomposition is in Table 42.

Any channel that aligns destinations with $\hat{\alpha}_j$ must run through informal reputation, because no centralised assessment ranks primary schools at this margin and the wedge is not published. What carries the signal is the transition histories of older cohorts, neighbour and sibling networks, and the standing a school holds locally, all tracking a quantity no published statistic reports.

Sorting intensity ρ implied by the destination Grade 8 estimate

The conditional estimate $\beta_{\alpha}^{G8} \approx 0.17$ from the destination Grade 8 revealed-preference equation (the wedge’s coefficient on the destination Grade 8 peer-mark outcome) supplies the empirically anchored sorting intensity on the standardised metric, yielding $\rho_{RP} \approx 0.17$ on the published construction, an upper anchor since it contains the same-feeder component of the peer mean (Appendix M), and $\rho_{RP} \approx 0.14$ on the re-standardised cross-feeder metric (the two anchors sit on slightly different standardised metrics). The Monte Carlo of §5.1 of the main text evaluates sorting intensity on the discrete grid $\rho \in \{0, 0.3, 0.6, 1.0\}$ and identifies $\rho = 0.6$ as an empirically plausible upper bound for the sorting-induced level distortion; the calibration’s rank-stability holds across the full grid. Both data-implied anchors sit below the lowest non-zero grid point ($\rho = 0.3$), so the empirical sorting intensity falls within the lower, empirically plausible portion of the Monte Carlo grid. The robustness claim for ranks in §5.1 of the main text is therefore unchanged by the empirical anchor; the empirical anchor additionally confirms that level interpretation is not at risk in the regime the data actually occupy. One disclosure is owed on the bound’s logic: the exercise is partly self-referential, in that it uses the calibration’s own $\hat{\alpha}_j$ to bound the sorting threat *to* the calibration’s $\hat{\alpha}_j$, so the bound is informative only if the calibration is approximately correct in the first place; we disclose this rather than treat the anchor as an external check.

The choice of quality signal therefore decides which sorting margin a study identifies. A school-choice analysis that proxies primary-school quality with raw Grade 7 marks, a common substitute where no calibrated measure exists, recovers the like-with-like matching of marks and no signal about the achievement of the cohort a student joins; on the intake margin it recovers the opposite sign to the wedge. This is a matter of direction, not precision: where calibrated quality and raw marks are negatively correlated at the school level, as they are here, a raw-mark proxy can report sorting away from quality exactly where households sort towards it (Hsieh and Urquiola 2006; Pop-Eleches and Urquiola 2013). The currency in which a researcher prices school quality sets the sign of the answer.

Caveats on movers and geographic resolution

Two named caveats supplement the three caveats in §6.1 of the main text.

Restriction to movers. The sample is restricted to students who attend a different Grade 7 and Grade 8 school (roughly half of linked students). If the implicit sorting on $\hat{\alpha}_j$ revealed by stayers exceeds the explicit sorting revealed by movers, the mover-only estimate understates population sorting intensity; in the reverse case, it overstates. The bipartite-network design of §6.1 of the main text requires within-feeder destination variation, which only movers supply; there is no direct test of stayer sorting on the available data. We report this as a named threat rather than a quantified correction.

Geographic resolution. The estimate is identified within province, within cohort, after partialling out origin quintile, urban-rural status, and no-fee status. The ρ parameter of §5.1 of the main text is silent on the cross-level distinction (within- vs across-province sorting); the empirical anchor therefore applies to within-province sorting, which is the empirically dominant geographic margin in the linked panel (1.9% of linked students change province in the Grade 7-to-Grade 8 window, and the estimation panel excludes them by construction).

School geography: a bound on the proximity-clustering channel

The revealed-preference design cannot, on the available data, separate *active* sorting on $\hat{\alpha}_j$, households choosing a destination because of its calibrated quality, from *passive* residential clustering, in which families relocate for reasons unrelated to the school signal and then enrol at whatever school is nearest. No student home location or family identifier exists in the source records (the raw extract is keyed only on student and school), so the four household-anchored designs that would identify the mechanism (home-to-school distance, equidistant-household choice sets, residential mover/stayer comparisons, and sibling fixed effects) are all infeasible. This appendix reports what the *school* coordinates alone can do: bound the contribution of one specific non-choice story, mechanical enrolment at the nearest school, by re-estimating the headline on origin-to-destination geographic cuts. It is a bound on a channel, **not** an identification of the mechanism.

The masterlist geocodes both the origin and destination school for 97.3 per cent of movers. The un-restricted row of Table 43 reproduces the published headline exactly (+0.165 on the destination Grade 8 margin, −0.069 on the Grade 7 margin), and restricting to the geo-complete sub-sample barely moves it (+0.155 on Grade 8); coverage attrition is not itself selecting on the result.

Table 43: Origin-to-destination geographic cuts of the revealed-preference estimate. Each cell is the coefficient on the origin school’s wedge (standardised) from the headline specification, re-estimated on the indicated sub-sample (or, for the interaction row, the coefficient on its product with distance); two-way clustered standard errors (origin, destination) in parentheses. Distance is the great-circle origin-to-destination school distance. Coefficients significant at the 5, 1 and 0.1 per cent levels carry one, two and three asterisks respectively. Sample: movers with both origin and destination schools geocoded, $n = 215,209$ (97.3 per cent of the headline sample); mathematics; Gauteng core, pooled 2018 and 2022 cohorts.

Specification	n	Dest. G7 peer (z)	Dest. G8 peer (z)
Headline (no geographic restriction)	221,093	−0.069* (0.030)	+0.165*** (0.018)
Geo-complete baseline	215,209	−0.082** (0.030)	+0.155*** (0.018)
Distance tertile 1 (nearest)	71,737	−0.288*** (0.042)	+0.138*** (0.028)
Distance tertile 2 (middle)	71,736	−0.052 (0.036)	+0.170*** (0.026)
Distance tertile 3 (farthest)	71,736	+0.063 (0.033)	+0.145*** (0.021)
$\hat{\alpha}_j \times$ distance (per +100 km)	215,209	+0.650*** (0.084)	−0.203*** (0.056)
Marginal $\hat{\alpha}_j$ at median distance (1.7 km)	215,209	−0.103*** (0.030)	+0.161*** (0.019)
Marginal $\hat{\alpha}_j$ at p90 distance (9.5 km)	215,209	−0.052 (0.029)	+0.145*** (0.018)
Same ward	77,354	−0.222*** (0.041)	+0.167*** (0.029)
Cross ward	137,855	−0.011 (0.030)	+0.143*** (0.019)

(continued)

Specification	n	Dest. G7 peer (z)	Dest. G8 peer (z)
Same municipality	202,853	-0.097** (0.031)	+0.159*** (0.019)
Cross municipality	12,356	+0.159*** (0.029)	+0.082** (0.027)

The calibrated destination Grade 8 loading survives every cut (Table 43). It weakens with distance but stays positive across the observed range: the $\hat{\alpha}_j \times$ distance interaction is -0.203 per $+100$ km ($p < 0.001$), which takes the marginal loading from $+0.161$ at the median origin-to-destination distance (1.7 km) to $+0.145$ at the 90th percentile (9.5 km), about 10 per cent lower, while the across-tertile estimates ($+0.138$ nearest, $+0.145$ farthest) move little. It survives the within- versus across-boundary contrast that most directly isolates the nearest-school story: among the 64.1 per cent of movers who cross a ward boundary the loading is $+0.143$ ($t = 7.7$), and among the 5.7 per cent who cross a *municipal* boundary, and who cannot plausibly be enrolling at the school nearest a within-ward home, it is $+0.082$ ($t = 3.0$), still about 53 per cent of the geo-complete baseline and decisively non-zero. A pure mechanical-proximity account predicts the cross-municipality cell should collapse; it does not.

The interpretive payoff is narrow and must be stated as such. What these cuts remove is the *vacuous* reading: that the positive Grade 8 result is an artefact of families simply attending the closest school, which happens to be a good one. They do **not** point towards active choice and away from clustering, because the non-choice story that survives is not nearest-school enrolment but **residential relocation into better-served areas**: a household that moves across a municipal boundary into a higher-quality neighbourhood and then enrolls locally produces exactly the long-displacement, boundary-crossing pattern just described, and that pattern is *strongest* in precisely the cross-municipality cell where the proximity bound looks most convincing. Residential moves are unobserved, so this channel cannot be excluded. The bound therefore tightens the *clustering* family of explanations: it kills mechanical proximity, leaving relocation, without separating clustering from choice. The non-separability caveat of §6.1 of the main text stands: the central revealed-preference claim is the equilibrium one, that raw marks identify the *wrong* sorting margin while the calibrated level identifies the right one, and it holds whether the alignment is produced by active choice or by sorting on a correlate of $\hat{\alpha}_j$.

One further reading cuts gently against pure relocation, and is reported with its own caveat. If long-distance switching were driven by households moving up an amenity or socioeconomic gradient, the destination Grade 7 *intake* margin would be expected to load *positively* on $\hat{\alpha}_j$ (richer catchments, higher-attaining intakes). The observed bottom-left coefficient is instead negative on the value-added margin, and, like the rest of the negative Grade 7 signature, it is **locally concentrated**: strongest among the nearest movers (distance tertile 1, -0.288) and decaying to insignificance among the farthest (tertile 3, 0.063 , $p = 0.055$), and it changes sign between within-municipality (-0.097) and cross-municipality ($+0.159$) movers. The negative intake loading is thus a short-range phenomenon, consistent with the bridge’s mechanical reading of the Grade 7 cell (§3 of the main text) rather than with a relocation-up-the-gradient story; but because it is local, it is at most a weak discriminator, and we do not lean on it.

Published matric: an own-feeder-immune corroboration

§6.1.6 of the main text re-runs the single-predictor test with the one academic quality signal published at the Grade 7 to Grade 8 choice margin: the destination secondary school’s published National Senior Certificate (NSC) Mathematics mean, matched on destination EMIS and standardised within cohort. The construction is the headline construction with only the outcome swapped: the same mover sample, the same seven controls, the same province-by-cohort fixed effects, and the same two-way clustering at origin and destination. The panel reproduces the two headline cells before any matric cell is read. Year alignment follows what a family could observe at the choice point: the 2018 Grade 8 cohort is matched to the NSC published over 2015 to 2017 (the record available when it chose, in late 2017) and the 2022 cohort to 2019 to 2021. The destination’s eventual NSC 2022 mean, four years past Grade 8 entry for the 2018 cohort, is reported as a forward check.

Table 44 collects the single-predictor coefficients across cohorts and outcome variants. The calibrated standing predicts the destination’s published matric in every cell, exceeds raw Grade 7 selectivity by roughly 0.144 to 0.049 on the pooled outcome (about 3 to one), and the two pooled intervals are disjoint ([0.104, 0.184] against [0.019, 0.078]). Two features set this corroboration apart from the headline. First, the outcome is own-feeder-immune: a published matric mean is built from cohorts that left before the choosing student arrived, so it cannot contain her own feeder-mates, and none of the leave-own-feeder-out machinery of Appendix M is needed to defend it. Second, the published matric is external to the calibration and so carries no construction overlap with $\hat{\alpha}_j$; the joint specification that reverses the calibrated coefficient on the destination peer-Grade-8 outcome (Appendix M) leaves it positive here, at +0.248 ($t = 6.0$).

Table 44: Origin-school signals as predictors of the destination secondary school’s published NSC mean, by cohort and outcome variant. Each cell is the coefficient (standard error) on the standardised origin signal from a separate single-predictor fit, with own Grade 7 mark and the seven controls partialled out; the outcome is standardised within cohort. Controls, province-by-cohort fixed effects and two-way clustering at origin and destination school follow the headline revealed-preference specification. Asterisks mark significance at the 5, 1 and 0.1 per cent levels. Sample: revealed-preference mover sample; Mathematics; Gauteng core; 2018 and 2022 cohorts.

Cohort / outcome	n	Wedge	Raw Grade 7	Raw Grade 8
2018 cohort, Maths NSC at choice (2015-17)	89,025	+0.143*** (0.021)	+0.031 (0.017)	+0.251*** (0.034)
2022 cohort, Maths NSC at choice (2019-21)	120,641	+0.163*** (0.026)	+0.052** (0.018)	+0.215*** (0.031)
Pooled, Maths NSC at choice	209,666	+0.144*** (0.020)	+0.049** (0.015)	+0.224*** (0.027)
2018 cohort, overall NSC at choice (2015-17)	89,192	+0.183*** (0.022)	+0.029 (0.017)	+0.312*** (0.031)
2018 cohort, Maths NSC 2022 (forward check)	91,657	+0.134*** (0.026)	+0.064** (0.020)	+0.294*** (0.038)

Appendix N: Decomposing $\hat{\beta}_1 = \lambda\rho$: persistence, reliability and horizon tests

Overview

The latent-variable derivation in Appendix O establishes that the Grade 7-to-Grade 8 persistence parameter converges in probability to the product $\lambda\rho$, where $\lambda \in (0, 1]$ is the persistence of latent learning across the Grade 7-to-Grade 8 transition and $\rho \in (0, 1]$ is the within-cell reliability (signal share) of the Grade 7 mark. The headline OLS estimate, $\hat{\beta}_1 = 0.601$ (2018 cohort) and 0.592 (2022 cohort), is therefore a product of two distinct factors that the baseline regression cannot separate. The purpose of this appendix is to decompose $\hat{\beta}_1$ into its two components using three independent tests, and to reconcile the tests into a single set of bounds on λ and ρ .

One-sided bounds without any test. From $\lambda\rho \approx 0.60$ and both factors bounded in $(0, 1]$, each factor satisfies $\lambda \geq 0.60$ and $\rho \geq 0.60$ regardless of which story, measurement error or genuine fade, dominates. This bound is sufficient for the paper’s substantive conclusions: the latent is at least 60% persistent and the Grade 7 mark is at least 60% reliable signal. The three tests in this appendix sharpen these bounds into a narrower interval.

Tests run. Three tests prescribed by the latent-variable framework are implemented in sequence:

1. **Test 1 (IV).** Instrument the Grade 7 Mathematics mark with the student’s Grade 7 English FAL mark. Under classical measurement error, the IV slope identifies $\lambda\rho_{\text{gen}}\rho_{\text{Maths}}^{-1} \approx \lambda$ (the persistence of a general academic latent). The IV jump relative to OLS isolates the attenuation component. Residual cross-subject heaping makes this a lower bound on λ rather than a clean point estimate.

2. **Test 2 (test–retest reliability).** Bound the within-cell reliability ρ of the Term 4 mark from pairs of the same student’s within-school-demeaned Grade 7 term marks that share no component. The Term 1 and Term 4 marks are not such a pair: in 2022 the Term 4 mark is in effect a year mark built from the earlier terms’ marks, and in both cohorts shared heaping at the 40% threshold ($6.09 \times$ its expected frequency in the joint $(T_1, T_4) = (40, 40)$ cell) adds to the overlap. Terms 1, 2 and 3 do not contain one another.
3. **Test 3 (horizon variation).** Re-estimate $\hat{\beta}_1$ using Grade 8 Term 3 and Term 4 marks as alternative endlines. If $\lambda < 1$, the slope should decline as the Gr7-to-Gr8 gap widens while ρ , a property of the Gr7 mark, remains roughly constant. A qualification applies: Grade 8 Term 4 is the year-end examination rather than a continuous-assessment term, so it may carry a higher ρ than Term 2 or Term 3; this limits interpretation of the T4 result as a clean horizon test.

Test 1: IV decomposition

The instrumental-variables estimator uses the Grade 7 English FAL Term 4 mark as an instrument for the Grade 7 Mathematics Term 4 mark, within the same origin-school $\times$ destination-grade fixed-effects structure as the headline OLS regression. The identifying assumption is that the measurement errors in the two subjects’ marks are uncorrelated conditional on the school FE: $\text{Cov}(\tilde{\varepsilon}_{\text{Eng}}^7, \tilde{\varepsilon}_{\text{Maths}}^7) = 0$ after within-cell demeaning. The student-level cross-subject heaping correlation at mark 40 is $r \approx 0.05\text{--}0.10$, substantially weaker than the within-subject cross-term correlation of $r = 0.27$ that disqualified same-subject earlier-term marks (Appendix B.6). The residual contamination biases $\hat{\beta}_1^{IV}$ downward towards OLS rather than upward, so the IV result is conservative: the true λ is at least as large as $\hat{\beta}_1^{IV}$.

Table 45: IV decomposition on one scale: instrument Grade 7 Mathematics with Grade 7 English FAL, both z-scored within province, cohort and subject. OLS and IV slopes on the same students, implied $\rho = \hat{\beta}_1^{OLS} / \hat{\beta}_1^{IV}$, and implied $\lambda \approx \hat{\beta}_1^{IV}$, per province and cohort year.

Province	Cohort	n	$\hat{\beta}_1^{OLS}$	$\hat{\beta}_1^{IV}$	SE(IV)	F (1st stage)	Implied ρ	Implied λ
EC	2018	92,774	0.464	0.740	0.0058	28,578	0.628	0.740
EC	2022	100,248	0.507	0.669	0.0043	57,859	0.758	0.669
GT	2018	63,729	0.580	0.760	0.0053	34,662	0.763	0.760
GT	2022	81,910	0.576	0.734	0.0042	64,479	0.784	0.734
LP	2018	84,905	0.519	0.721	0.0047	42,908	0.720	0.721
LP	2022	110,641	0.574	0.732	0.0036	96,149	0.784	0.732
MP	2018	56,795	0.569	0.771	0.0059	28,651	0.738	0.771
MP	2022	66,329	0.594	0.759	0.0048	58,879	0.783	0.759
NW	2018	37,717	0.582	0.838	0.0078	17,098	0.694	0.838
NW	2022	49,717	0.602	0.793	0.0057	33,526	0.759	0.793
ZN	2018	97,324	0.489	0.712	0.0048	48,019	0.686	0.712
ZN	2022	132,883	0.573	0.762	0.0035	105,118	0.752	0.762

Instrument: Grade 7 English FAL Term 4 mark for the same student. *Scale:* both marks are z-scored within province $\times$ cohort $\times$ subject, as in the headline regression, so both slopes are in SD of the Grade 8 mark per SD of the Grade 7 mark, and implied ρ is the ratio of two slopes fitted on the same students. *Identifying assumption:* $\text{Cov}(\tilde{\varepsilon}_{\text{Eng}}^7, \tilde{\varepsilon}_{\text{Maths}}^7) = 0$ after within-cell demeaning. *Residual cross-subject heaping biases $\hat{\beta}_1^{IV}$ downward, so implied ρ is an upper bound and implied λ is a lower bound.* *Sample:* National six-province Mathematics panel, inner join on Mathematics $\cap$ English FAL student IDs.

With both marks z-scored within province, cohort and subject, as in the headline regression, the n-weighted pooled IV slope is $\hat{\beta}_1^{IV} = 0.746$ (2018) and 0.737 (2022), against OLS slopes of 0.522 and 0.567 on the same students. Those OLS slopes sit below the Gauteng headline of 0.601 and 0.592 because the sample differs: it spans six provinces and keeps only students with an English FAL mark, and in Gauteng alone those students

give 0.580 and 0.576. First-stage F -statistics range from 17,098 to 105,118 across the 12 province-by-cohort cells, far above any conventional weak-instrument threshold. The implied reliability, OLS over IV on the same students, is $\rho \approx 0.699$ (2018) and 0.769 (2022). Measurement-error attenuation (Story 1: $\rho < 1$) is therefore a large part of the gap between $\hat{\beta}_1$ and one, but not all of it: the IV slope itself, a lower bound on λ , is 0.746 and 0.737, so genuine fade (Story 2: $\lambda < 1$) is of a similar order. Attenuation is the larger factor in 2018, and the two are about equal in 2022 (Appendix N). Because the cross-subject heaping biases $\hat{\beta}_1^{IV}$ downward (towards OLS), the denominator of $\hat{\rho}^{IV} = \hat{\beta}_1^{OLS} / \hat{\beta}_1^{IV}$ is too small, and $\hat{\rho}^{IV}$ is consequently biased upward. The IV-implied ρ values are therefore upper bounds on the true ρ .

A note on the sign of the IV bias and its implication: the cross-subject heaping generates positively correlated measurement errors ($r \approx 0.05$ – 0.10), which means the instrument is not fully orthogonal to the error in the Mathematics mark, and TSLS will be biased towards the OLS estimate. With the IV biased towards OLS (i.e., downward relative to the true λ), the observed $\hat{\beta}_1^{IV} \approx 0.75$ in 2018 already substantially corrects the attenuation: the true λ is at least 0.75 (2018) and 0.74 (2022).

Test 2: Direct reliability estimation

The within-cell reliability ρ is defined as the signal share of individual Grade 7 mark variance within an origin-school cell:

$$\rho = \frac{\text{Var}(\tilde{a}^7)}{\text{Var}(\tilde{a}^7) + \text{Var}(\tilde{\varepsilon}^7)},$$

where tildes denote within-origin-school-demeaned quantities. This is a student-level signal-to-total ratio, distinct from the school-level Kane–Staiger reliability $\rho^{KS} \approx 0.97$ estimated in §4.2 of the main text. The Kane–Staiger object measures how reliably the estimated school FE tracks the true school effect; it is large because many students per school reduce estimation noise. The student-level ρ measures how much of a single student’s single-sitting mark variance is true learning signal, a different question and a different quantity.

The test–retest estimator correlates two within-school-demeaned marks from the same student within the same academic year, treating the across-term correlation as an estimate of the signal share under the assumption that the latent a^7 is stable across terms. Table 46 reports the two pairs first used, Term 1 against Term 4 ($\hat{\rho}_{T1,T4}$) and Term 3 against Term 4 ($\hat{\rho}_{T3,T4}$). Neither is a retest. In 2022 the Term 4 mark is in effect a year mark: within a school the Term 1 to Term 3 marks explain 94% of its variance (71% in 2018), and the one-factor loading it implies has square 1.14, above the ceiling of one that holds for a mark sharing no error with the others (0.83 in 2018). Both pairs also share heaping: the joint $(T_1, T_4) = (40, 40)$ cell is $6.09\times$ its expected frequency under independence. The restated test uses only the three pairs among Terms 1 to 3.

Table 46: Test-retest reliability estimates per province and cohort year. $\hat{\rho}_{T1,T4}$ and $\hat{\rho}_{T3,T4}$ are upper bounds on the true ρ due to shared heaping. Implied $\lambda = \hat{\beta}_1 / \hat{\rho}$ uses the headline OLS slope. IV columns reproduced cell by cell from Table 45 for comparison.

Province	Cohort	$\hat{\rho}_{T1,T4}$	$\hat{\lambda}_{T1,T4}$	$\hat{\rho}_{T3,T4}$	$\hat{\lambda}_{T3,T4}$	IV $\hat{\rho}$	IV $\hat{\lambda}$
EC	2018	0.651	0.924	0.702	0.857	0.628	0.740
EC	2022	0.817	0.724	0.800	0.740	0.758	0.669
GT	2018	0.750	0.801	0.770	0.781	0.763	0.760
GT	2022	0.860	0.688	0.824	0.718	0.784	0.734
ZN	2018	0.713	0.842	0.744	0.808	0.686	0.712
ZN	2022	0.834	0.710	0.807	0.734	0.752	0.762
LP	2018	0.709	0.848	0.743	0.809	0.720	0.721
LP	2022	0.834	0.710	0.810	0.731	0.784	0.732
MP	2018	0.709	0.848	0.730	0.824	0.738	0.771
MP	2022	0.837	0.708	0.787	0.753	0.783	0.759

(continued)

Province	Cohort	$\hat{\rho}_{T1,T4}$	$\hat{\lambda}_{T1,T4}$	$\hat{\rho}_{T3,T4}$	$\hat{\lambda}_{T3,T4}$	IV $\hat{\rho}$	IV $\hat{\lambda}$
NW	2018	0.718	0.837	0.719	0.836	0.694	0.838
NW	2022	0.846	0.700	0.805	0.735	0.759	0.793

Method: Pearson correlation of within-origin-school-demeaned Grade 7 Term 1 (or Term 3) mark against Grade 7 Term 4 mark, for the same student. Schools with fewer than 5 students dropped. Implied $\hat{\lambda} = \hat{\beta}_1^{OLS} / \hat{\rho}$, using $\hat{\beta}_1^{OLS} = 0.601$ (2018) and 0.592 (2022). Neither $\hat{\rho}_{T1,T4}$ nor $\hat{\rho}_{T3,T4}$ is a retest: in 2022 the Term 4 mark is built largely from the earlier terms' marks, and both pairs share heaping at 40. The restated bound, from pairs among Terms 1 to 3, is in the text. IV $\hat{\rho}$ is an upper bound (see text); the IV columns are the same cell's values from Table 45, on one scale. Sample: National six-province Mathematics panel, intersection of Grade 7 T1 (or T3) and T4 mark availability.

The n-weighted pooled results are: $\hat{\rho}_{T1,T4} = 0.707$ (2018) and 0.837 (2022); $\hat{\rho}_{T3,T4} = 0.736$ (2018) and 0.807 (2022). Pooled, these exceed the IV-implied ρ of 0.699 (2018) and 0.769 (2022), the direction heaping bias predicts, though in 2018 only narrowly. Cell by cell, the Term 1 with Term 4 estimate exceeds the IV-implied ρ in 9 of the 12 province-by-cohort cells of Table 46, and the Term 3 with Term 4 estimate in 11. Every exception is a 2018 cell: Gauteng, Limpopo and Mpumalanga for Term 1, and Mpumalanga for Term 3. The three pairs among Terms 1 to 3 give each term's loading on the signal that is stable within the year; projecting the Term 4 mark on those terms then bounds its signal share at [0.605, 0.966] (2018) and [0.774, 0.968] (2022). The lower end assumes only that the part of the Term 4 mark the earlier terms do not predict does not load negatively on learning. One term's mark is less reliable, at 0.654 (2018) and 0.609 (2022).

The Term 1 against Term 4 correlation rises from 0.71 in 2018 to 0.84 in 2022. The IV-implied ρ , which heaping shared across terms of the same subject does not touch, rises too, from 0.699 to 0.769, about half the test-retest rise of 0.13. How the Term 4 mark is built explains the rest of the rise, not heaping. The Term 1 to Term 3 marks explain 71% of the Term 4 mark's within-school variance in 2018 and 94% in 2022; the break falls between 2019 and 2021, in all six provinces and in Grade 8 as well. A Term 1 mark that is part of the Term 4 mark does not retest it.

Test 3: Horizon variation

The theoretical prediction is that $\hat{\beta}_1^{(T_k)} \approx \lambda^{(T_k)} \rho$, where $\lambda^{(T_k)}$ is the persistence from Grade 7 Term 4 to Grade 8 Term k . If persistence genuinely decays ($\lambda < 1$), a longer horizon should yield a lower slope while ρ , a property of the Grade 7 mark, remains constant. This test uses Grade 8 Term 3 as a mid-year endline and Grade 8 Term 4 as a year-end endline, comparing both against the Term 2 baseline used in the headline regression. All marks are standardised within province $\times$ cohort $\times$ term to enable direct comparison of slopes across endline terms.

An important qualification: Grade 8 Term 4 is the year-end examination in the South African school calendar. It is a summative assessment that captures cumulative within-year learning and may therefore have a higher per-occasion reliability ρ than the continuous-assessment Term 2 or Term 3 marks. If ρ is higher at Term 4, the term-by-term slope comparison does not cleanly isolate λ , because $\hat{\beta}_1^{(T_4)} = \lambda^{(T_4)} \rho^{(T_4)}$ may exceed $\hat{\beta}_1^{(T_2)} = \lambda^{(T_2)} \rho^{(T_2)}$ even if λ decays, because the ρ increase more than offsets the λ decline. This confound limits interpretation of the T4 result.

Table 47: Horizon variation: TWFE slope $\hat{\beta}_1$ at Grade 8 Terms 2, 3, and 4 as alternative endlines. Standardised marks, n-weighted pooled across six provinces, by cohort. Δ from T2 is the change relative to the Term 2 baseline.

Cohort	Term	$\hat{\beta}_1$	OLS baseline (T2)	Δ from T2
2018	2	0.550	0.601	—

(continued)

Cohort	Term	$\hat{\beta}_1$	OLS baseline (T2)	Δ from T2
2018	3	0.538	0.601	-0.012
2018	4	0.601	0.601	+0.050
2022	2	0.580	0.592	—
2022	3	0.537	0.592	-0.042
2022	4	0.654	0.592	+0.074

Method: *TWFE feols()* with origin-school $\times$ destination-grade FE, same specification as headline OLS. Marks z-scored within province $\times$ cohort $\times$ term. Each term's column uses every student with a valid Grade 8 mark in that term, so the samples differ slightly across terms. The OLS column repeats the Gauteng headline, $\hat{\beta}_1 = 0.601/0.592$. Δ is each term's pooled slope minus this table's Term 2 slope. Sample: National six-province Mathematics panel, 2018 and 2022 cohorts, using Grade 8 Terms 3 and 4 marks.

The pattern is a V-shape: $\hat{\beta}_1^{(T_3)} < \hat{\beta}_1^{(T_2)} < \hat{\beta}_1^{(T_4)}$. The Term 3 slope is below Term 2 in both cohorts ($\Delta = -0.012$ in 2018 and -0.042 in 2022), consistent with the λ -decay prediction at a roughly one-term longer horizon. The Term 4 slope is above Term 2 ($\Delta = +0.050$ in 2018 and $+0.074$ in 2022). As discussed, the Term 4 rise is not consistent with simple λ decay; it is most plausibly explained by Term 4 being the year-end examination with a higher per-occasion ρ . The within-SBA T3 result provides the cleanest horizon test and is consistent with $\lambda < 1$.

Synthesis: bounds on λ and ρ

The three tests are over-identified. Each delivers a bound rather than a clean point estimate because each test's source of bias runs in a known direction:

- **IV lower bound on λ :** residual cross-subject heaping biases $\hat{\beta}_1^{IV}$ downward towards OLS $\rightarrow$ IV identifies $\lambda \geq 0.746$ (2018) / 0.737 (2022).
- **Test–retest bound on ρ :** no Grade 7 term mark retests the Term 4 mark without sharing a component with it, so the bound is built from the pairs among Terms 1 to 3 $\rightarrow \rho \leq 0.966$ (2018) / 0.968 (2022).
- **IV upper bound on ρ :** from $\hat{\rho}^{IV} = \hat{\beta}_1^{OLS} / \hat{\beta}_1^{IV}$, and with $\hat{\beta}_1^{IV}$ biased towards OLS (downward), $\hat{\rho}^{IV}$ is biased *upward*, so it is also an upper bound on true ρ : $\rho \leq 0.699$ (2018) / 0.769 (2022).
- **Horizon qualitative confirmation:** T3 result ($\Delta < 0$) confirms $\lambda < 1$ independently of the measurement approach.

Combining the IV and test–retest bounds:

$$\rho \leq 0.699 \quad (2018), \quad \rho \leq 0.769 \quad (2022) \quad (\text{IV}); \quad \rho \leq 0.966, 0.968 \quad (\text{test–retest}).$$

Because both the IV and the test–retest estimates of ρ are upper bounds, the lower end of each interval ($0.699 / 0.769$) is the more credible anchor for the true ρ , with the upper end indicating the maximum that measurement artefact could allow. Two lower bounds on λ are available: the IV slope, $\lambda \geq 0.746$ (2018) and 0.737 (2022), and the test–retest-implied lower bound, $\lambda \geq 0.622$ (2018) and 0.612 (2022). The test–retest floor sits barely above the slope itself, so the IV figure carries the decomposition. In 2022 the two legs also disagree: the Term 4 mark's stable share is at least 0.774 , above the IV-implied ρ , so part of what is stable across the year's terms is not shared with the English mark. Taking the test–retest floor as the conservative bound:

$$\lambda \geq 0.622 \quad (2018), \quad \lambda \geq 0.612 \quad (2022).$$

Reconciliation of IV-implied and TR-implied λ . Both methods deliver lower bounds on λ ; neither delivers an upper bound. The expected bias structure puts the IV above the test–retest figure: the IV is

biased towards OLS by a small amount (cross-subject heaping $r \approx 0.05\text{--}0.10$), so $\hat{\beta}_1^{IV}$ is only modestly below the true λ , while the test–retest bound is lower because it rests on the Term 4 mark’s stable share within the year, which no Grade 7 retest can pin down more tightly once the Term 4 mark contains the earlier terms. The two are not computed on the same students. The test–retest λ divides the Gauteng headline slope by a six-province reliability; the IV slope and its ρ come from the six-province students with an English FAL mark, whose OLS slope is lower (0.522 against 0.601 in 2018). Both methods place λ below one. The IV puts it near 0.746; the test–retest bound only rules out values below 0.622 (2018) and 0.612 (2022).

Which story dominates. On the same students, the IV jump runs from OLS ≈ 0.52 to IV ≈ 0.75 in 2018 and from 0.57 to 0.74 in 2022. Measurement-error attenuation ($\rho < 1$, Story 1) is the larger part of the gap between $\hat{\beta}_1$ and one in 2018; in 2022 attenuation and fade are about equal. Since both estimates of ρ are upper bounds, the Grade 7 mark carries at most about 70% (2018) and 77% (2022) true learning signal within the within-school cell. Genuine persistence fade ($\lambda < 1$, Story 2) is present in both cohorts, as independently confirmed by the T3 horizon result: on the IV estimate about 25% (2018) and 26% (2022) of true learning fades across the six-month Grade 7-to-Grade 8 transition. The test–retest bound cannot narrow this. Heaping in the assessments is, ironically, the reason the exact split cannot be pinned down more precisely: heaping contaminates both the IV exclusion restriction and the test–retest estimator, in each case in the same direction (downward towards OLS / upward towards one). The two contaminations jointly bracket the true values from above (test–retest ρ , IV λ) without providing a clean lower bound on λ other than the universal $\lambda \geq 0.60$.

Comparison with the existing literature. The IV-corrected persistence of $\lambda \approx 0.75$ (2018) and 0.74 (2022) for a six-month horizon sits above the persistence rates reported in the value-added literature at annual horizons. Kane and Staiger (2008), in an experimental evaluation of teacher effects in Los Angeles, find that teacher impacts fade out by roughly 50 per cent per year, an annual persistence of ≈ 0.5 , the lower end of the commonly cited 0.5–0.9 range. A higher λ at our shorter six-month horizon is expected, because less time has elapsed for learning to fade. Andrabi, Das, Khwaja and Zajonc (2011, *Do Value-Added Estimates Add Value? Accounting for Learning Dynamics*, AEJ: Applied Economics; World Bank PRWP 5066, 2009), who address attenuation explicitly with dynamic-panel methods on Pakistani data, find that “only a fifth to a half of achievement persists between grades”, an attenuation-corrected persistence of roughly 0.2–0.5 over annual horizons;³ the gap between their figures and ours is consistent with a shorter horizon and a different institutional setting.

What the noisy Grade 7 mark does to the wedge

The noisy Grade 7 mark biases the wedge as well as the slope. Within a school, the regression discounts each student’s mark for its noise, which is why $\hat{\beta}_1$ converges to $\lambda\rho$. A school’s mean mark needs no such discount, because the noise averages out over its students, yet the wedge subtracts the discounted slope times that mean. To first order and net of destination effects (the caveat of [Appendix O.7](#) applies), the wedge therefore converges to

$$\hat{\alpha}_j \xrightarrow{p} \beta_1 s_j + (\lambda - \beta_1) \bar{a}_j^7 = \rho \lambda s_j + \lambda(1 - \rho) \bar{a}_j^7,$$

where $\bar{a}_j^7$ is the mean latent Grade 7 learning of the school’s students. The first term is the difficulty term λs_j of §3.1 of the main text, shrunk by the reliability. The second is an intake term: it makes the assessments of a school whose students learned less by the end of Grade 7 look easier than they are. With a perfectly reliable mark ($\rho = 1$) both distortions vanish.

How much this matters can be read off the data without fixing λ . At any slope b the wedge is exactly $\hat{\alpha}_j(b) = \text{FE}_j(\text{Grade 8}) - b \text{FE}_j(\text{Grade 7})$, with both feeder effects net of destination effects, and under the model $b = \lambda$ removes the intake term. Figure 25 recomputes how much easier no-fee primaries’ assessments

³In the working-paper version (World Bank PRWP 5066), the 0.2–0.5 range is the dynamic-panel estimate correcting for both measurement error and student heterogeneity; correcting for measurement error alone gives 0.7–0.8 in that paper. This is a distinct study from Andrabi, Das and Khwaja (2017), *Report Cards* (AER 107(6): 1535–1563), which concerns an information intervention rather than persistence estimation.

are at every slope from 0.50 to 1.10. The contrast rises with the slope, by 0.020 (2018) and 0.033 (2022) for each 0.1, because no-fee primaries send students into the same secondary schools with higher Grade 7 marks than fee-charging primaries do: higher by 0.20 SD (2018) and 0.33 SD (2022), net of destination. Under the model, their easier assessments lift those marks by more than their weaker intake lowers them. The contrast would reach zero only at a slope of -1.34 (2018) or -1.11 (2022). Any slope above the OLS value therefore widens the no-fee gap, whichever estimate of λ is right.

At the instrumental-variables slope on the headline sample, 0.760 (2018) and 0.734 (2022), the Gauteng row of Table 45 with reliabilities of 0.763 and 0.784, the pooled no-fee contrast is 0.509 (SE 0.024), against 0.470 (SE 0.021) at OLS. Across the instrumental-variables and test–retest slopes it runs from 0.509 to 0.515. The wedge’s gradient across school poverty quintiles moves the same way, from 0.124 to 0.133 SD per quintile. At the IV slope the intake term supplies 0.063 of the OLS contrast (13%); across the instrumental-variables and test–retest slopes it supplies 13–15% of the contrast and 9–10% of the wedge’s variance across schools. The shrinkage of the difficulty term hides 0.102. Attenuation hides more of the no-fee gap than intake adds, so the OLS figure understates the no-fee gradient.

The corrected wedge is more dispersed and ranks schools almost exactly as the published one does. Its SD is 0.554 against 0.471 (2018) and 0.527 against 0.468 (2022), and it correlates at 0.994 and 0.995 with the published wedge, Pearson and Spearman alike. Between the cohorts, over the 1,280 schools estimated in both, the corrected wedge correlates at 0.586 (precision-weighted, SE 0.029), against 0.624 for the published wedge: a difference of -0.039 , with a 95% interval from -0.047 to -0.029 . The intake term alone persists more strongly, at 0.717. About 0.04 of the published wedge’s persistence is therefore intake, and the difficulty part persists without it.

The sweep is algebra and needs no model. Reading $b = \lambda$ as the slope that removes the intake term rests on the model of Appendix O, and four features of the data qualify it.

First, the instrument’s sample. The panel holds no English Home Language mark, so the IV is fitted only on students with an English FAL mark: 60% of the headline sample’s students and 913 of its 1,424 feeders in 2018 (60% and 928 of 1,479 in 2022). They lean poor: 85% of them attend no-fee schools, against 61% of the headline sample (87% against 65% in 2022). The slope is applied to every feeder on the assumption that λ is common.

Second, correlated errors across subjects. Illness in an examination week, or heaping common to a school’s marks in both subjects, pulls the IV towards OLS, so the IV understates λ . The contrast rises with the slope, so this cuts towards a larger gradient, not a smaller one.

Third, general rather than mathematics-specific learning. English and mathematics share a general latent, so the IV slope is the persistence of that latent. It sits close to the test–retest values, and each of them gives a pooled contrast between 0.509 and 0.515.

Fourth, reliability that varies by school. The split assumes one reliability for every school. If no-fee primaries’ marks are noisier, the shrinkage differs by school and the split above does not capture it; this is not tested here. Appendix H.8 finds the Grade 7 slope flatter among no-fee primaries, which noisier no-fee marks would produce, among other channels.

Appendix O: Deriving the Estimating Equation from First Principles

This appendix carries the full derivation that §3.1 of the main text states. It starts from three structural primitives (a Grade 7 measurement equation, a Grade 8 measurement equation, and a transition law) and derives four reduced forms by varying which of the two structural mechanisms (measurement error; genuine fade) are active. It then works through the probability-limit algebra that yields $\hat{\beta}_1 \xrightarrow{p} \lambda\rho$, defines the continuum structure of $\hat{\alpha}_j$, and closes with a technical summary table. No data are used; the derivation is purely algebraic.

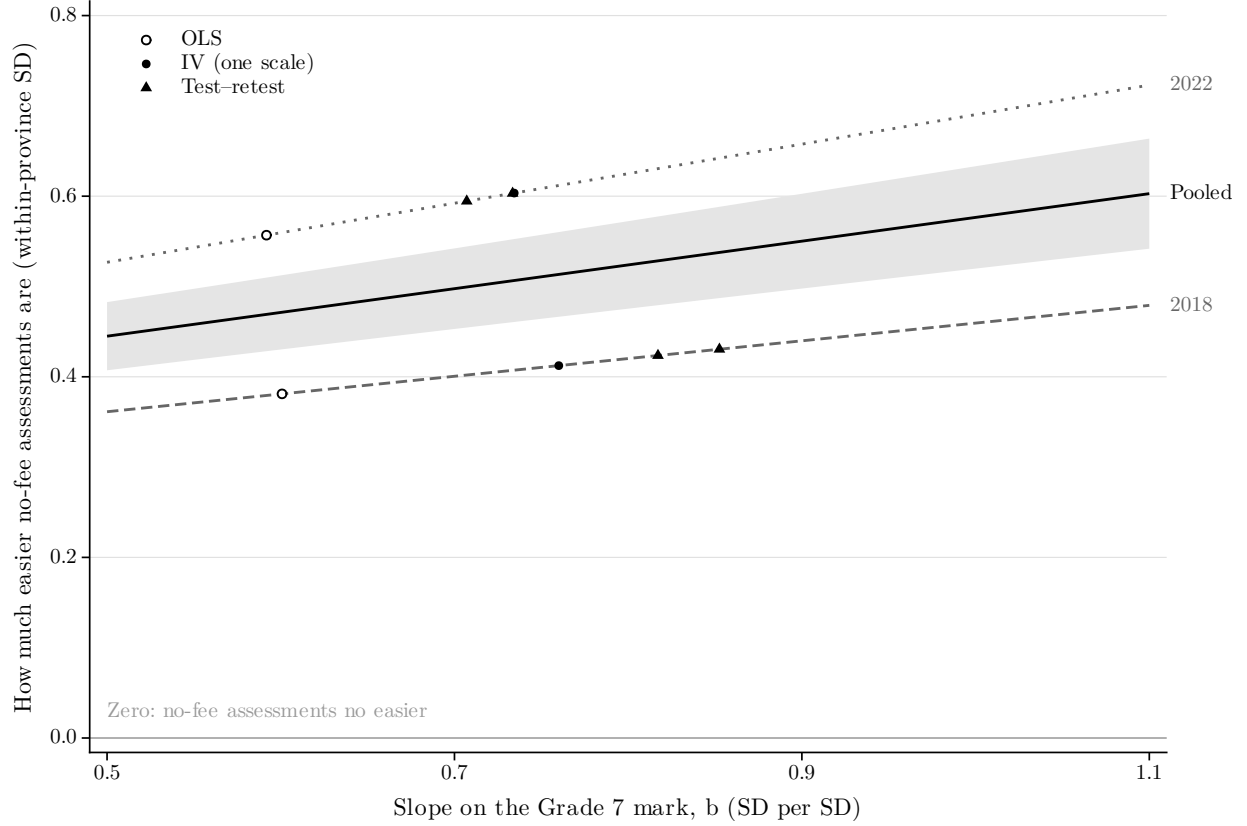


Figure 25: **Correcting for the noisy Grade 7 mark widens the no-fee gap.**

Note. Each line is how much easier no-fee primaries' assessments are: the precision-weighted difference in $-\hat{\alpha}_j(b)$ between no-fee and fee-charging primaries, positive when no-fee primaries' assessments are easier, with the wedge recomputed at each slope b on the Grade 7 mark as $\hat{\alpha}_j(b) = \text{FE}_j(\text{Grade 8}) - b \text{FE}_j(\text{Grade 7})$, both feeder effects net of destination effects. At fixed weights the contrast is exactly linear in b . Points mark the named slopes on each cohort's line: OLS, $\hat{\beta}_1$ (0.601 in 2018, 0.592 in 2022); the English FAL instrumental-variables slope on one scale (0.760 and 0.734); and the two test-retest values $\hat{\beta}_1/\hat{\rho}$. Pooled, the contrast is 0.470 (SE 0.021) at the OLS slope and 0.509 (SE 0.024) at the instrumental-variables slope. Contrasts are in within-province standard deviations of the Grade 8 mark; slopes are in SD of the Grade 8 mark per SD of the Grade 7 mark. The band is the pooled contrast's 95% interval, with standard errors clustered by school and b and the weights held fixed. No significance star appears in this figure; where stars appear elsewhere in this document they follow one ladder, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an em dash. None arises here. Source: analyses/wedge_intake_loading/outputs/tables/slope_sweep.csv. **Sample:** Gauteng core, 2018 and 2022 cohorts, the headline estimation sample; the contrast covers 2,871 school-cohorts with a fee status.

O.1 Notation

The model has one underlying object, a student's true learning, measured twice on scales that two different schools control.

Symbol	Meaning	Type
a_{ij}^7	latent (true) learning of student i at end of Grade 7, origin school j	unobserved
a_{ijk}^8	latent learning of student i at Grade 8 Term 2, destination cell k	unobserved
g_{ij}^7	observed Grade 7 mark (Gr7Term4)	observed
g_{ijk}^8	observed Grade 8 mark (Gr8Term2)	observed
s_j	origin mark-scale offset: school-level shift between true learning and the Grade 7 mark	unobserved, structural
d_k	destination mark-scale offset: cell-level shift between true learning and the Grade 8 mark	unobserved, structural
$\varepsilon_{ij}^7, \varepsilon_{ijk}^8$	idiosyncratic measurement error in each mark	unobserved, mean zero
c	average six-month learning gain (growth constant)	structural
ν_{ijk}	idiosyncratic deviation from average growth	unobserved, mean zero
λ	persistence of latent learning across the transition	structural, $\in (0, 1]$
ρ	reliability of the Grade 7 mark (signal share of within-cell variance)	structural, $\in (0, 1]$
β_1	slope on the lagged mark in the estimating regression	estimated
α_j, γ_k	origin and destination fixed effects in the estimating regression	estimated

Sign convention. A school with harder assessments marks below true learning:

$$g_{ij}^7 = a_{ij}^7 - s_j - \varepsilon_{ij}^7, \quad \text{equivalently} \quad a_{ij}^7 = g_{ij}^7 + s_j + \varepsilon_{ij}^7. \quad (\text{O.1})$$

Harder $\Rightarrow s_j > 0$; a hard destination paper $\Rightarrow d_k > 0$.

Classical assumptions. The latent a^7 , both measurement errors $\varepsilon^7, \varepsilon^8$, and the growth shock ν are mutually uncorrelated and each error has mean zero. These are the standard errors-in-variables assumptions; relaxing them is a robustness question addressed in [Appendix N](#).

O.2 The two measurement equations and the transition

Grade 7 measurement is equation (O.1) above. **Grade 8 measurement** relates the same student's Grade 8 latent to the observed anchor mark:

$$a_{ijk}^8 = g_{ijk}^8 + d_k + \varepsilon_{ijk}^8, \quad \text{equivalently} \quad g_{ijk}^8 = a_{ijk}^8 - d_k - \varepsilon_{ijk}^8. \quad (\text{O.2})$$

The destination offset d_k does not depend on origin j : the origin school neither sets nor grades the Grade 8 assessment, so equation (O.2) is origin-uncontaminated.

Transition law. Replacing level-preserving growth with partial persistence:

$$a_{ijk}^8 = c + \lambda a_{ij}^7 + \nu_{ijk}, \quad 0 < \lambda \leq 1. \quad (\text{O.3})$$

With $\lambda = 1$ this is level-preserving (no mean reversion); with $\lambda < 1$ a student's Grade 7 standing compresses by factor λ across the transition. The section below uses both the $\lambda = 1$ and $\lambda < 1$ versions.

O.3 Reduced form 1: benchmark ($\beta_1 = 1$)

Suppose marks are measured exactly ($\varepsilon^7 = \varepsilon^8 = 0$) and the transition is level-preserving ($\lambda = 1$).

Step A. Start from (O.2):

$$g_{ijk}^8 = a_{ijk}^8 - d_k.$$

Step B. Substitute the $\lambda = 1$ transition, $a_{ijk}^8 = a_{ij}^7 + c + \nu_{ijk}$:

$$g_{ijk}^8 = a_{ij}^7 + c + \nu_{ijk} - d_k.$$

Step C. Substitute (O.1) with $\varepsilon^7 = 0$, so $a_{ij}^7 = g_{ij}^7 + s_j$:

$$g_{ijk}^8 = (g_{ij}^7 + s_j) + c + \nu_{ijk} - d_k.$$

Step D. Collect by type:

$$g_{ijk}^8 = c + g_{ij}^7 + s_j - d_k + \nu_{ijk}.$$

Step E. Make the coefficient on g^7 explicit (it is exactly 1); name the error $\varepsilon_{ijk}^* \equiv \nu_{ijk}$:

$$\boxed{g_{ijk}^8 = c + 1 \cdot g_{ij}^7 + s_j - d_k + \varepsilon_{ijk}^*} \quad (\text{O.4})$$

Under (O.4) the origin fixed effect $\hat{\alpha}_j$ recovers s_j , pure difficulty. The benchmark prediction is $\beta_1 = 1$: relative standing passes through intact.

O.4 Reduced form 2: pure attenuation ($\beta_1 = \rho$)

Keep level-preserving growth ($\lambda = 1$) but now allow measurement error ($\varepsilon^7 \neq 0$, $\varepsilon^8 \neq 0$). Steps A–D follow equation (O.4) exactly, giving the same structural form, except the composite error now includes the measurement errors. The regressor $g_{ij}^7 = a_{ij}^7 - s_j - \varepsilon_{ij}^7$ carries $-\varepsilon^7$ in the regressor and the error $\varepsilon^* = \varepsilon^7 - \varepsilon^8 + \nu$ carries $+\varepsilon^7$. OLS sees a regressor negatively correlated with its error.

The within-transform (Frisch–Waugh–Lovell) annihilates origin and destination fixed effects because s_j is constant within origin j and d_k is constant within destination k . Writing $\tilde{x}$ for the within-cell deviation:

$$\tilde{g}_{ij}^7 = \tilde{a}_{ij}^7 - \tilde{\varepsilon}_{ij}^7, \quad \tilde{g}_{ijk}^8 = \tilde{g}_{ij}^7 + \tilde{\varepsilon}_{ijk}^*.$$

The OLS probability limit is

$$\hat{\beta}_1 \xrightarrow{p} \frac{\text{Cov}(\tilde{g}^7, \tilde{g}^8)}{\text{Var}(\tilde{g}^7)} = \frac{\text{Var}(\tilde{g}^7) + \text{Cov}(\tilde{g}^7, \tilde{\varepsilon}^*)}{\text{Var}(\tilde{g}^7)}.$$

Nine-covariance expansion of the numerator. Using $\tilde{g}^7 = \tilde{a}^7 - \tilde{\varepsilon}^7$ and $\tilde{\varepsilon}^* = \tilde{\varepsilon}^7 - \tilde{\varepsilon}^8 + \tilde{\nu}$, the cross-covariance $\text{Cov}(\tilde{g}^7, \tilde{\varepsilon}^*)$ expands into nine terms:

$$\text{Cov}(\tilde{a}^7 - \tilde{\varepsilon}^7, \tilde{\varepsilon}^7 - \tilde{\varepsilon}^8 + \tilde{\nu}).$$

By the classical assumptions every term is zero except $\text{Cov}(-\tilde{\varepsilon}^7, \tilde{\varepsilon}^7) = -\text{Var}(\tilde{\varepsilon}^7)$:

$$\underbrace{\text{Cov}(\tilde{a}^7, \tilde{\varepsilon}^7)}_0 - \underbrace{\text{Cov}(\tilde{a}^7, \tilde{\varepsilon}^8)}_0 + \underbrace{\text{Cov}(\tilde{a}^7, \tilde{\nu})}_0 - \underbrace{\text{Cov}(\tilde{\varepsilon}^7, \tilde{\varepsilon}^7)}_{\text{Var}(\tilde{\varepsilon}^7)} + \underbrace{\text{Cov}(\tilde{\varepsilon}^7, \tilde{\varepsilon}^8)}_0 - \underbrace{\text{Cov}(\tilde{\varepsilon}^7, \tilde{\nu})}_0 = -\text{Var}(\tilde{\varepsilon}^7).$$

So the numerator is $\text{Var}(\tilde{g}^7) - \text{Var}(\tilde{\varepsilon}^7)$. The denominator uses $\tilde{g}^7 = \tilde{a}^7 - \tilde{\varepsilon}^7$ with the two pieces uncorrelated:

$$\text{Var}(\tilde{g}^7) = \text{Var}(\tilde{a}^7) + \text{Var}(\tilde{\varepsilon}^7).$$

Combining:

$$\hat{\beta}_1 \xrightarrow{p} \frac{\text{Var}(\tilde{a}^7)}{\text{Var}(\tilde{a}^7) + \text{Var}(\tilde{\varepsilon}^7)} \equiv \rho. \quad (\text{O.5})$$

The reliability ρ is the signal share of the within-cell variance of the Grade 7 mark. Because variances are non-negative, $\rho \leq 1$: measurement error always pulls the slope towards zero. Under this pure-attenuation story the latent is fully persistent; the sub-unit slope is entirely a measurement artefact.

O.5 Reduced form 3: pure fade ($\beta_1 = \lambda$)

Now reverse the experiment: marks are exact ($\varepsilon^7 = \varepsilon^8 = 0$) but the transition has $\lambda < 1$.

Step A. Start from (O.2) with $\varepsilon^8 = 0$:

$$g_{ijk}^8 = a_{ijk}^8 - d_k.$$

Step B. Substitute the partial-persistence transition (O.3):

$$g_{ijk}^8 = (c + \lambda a_{ij}^7 + \nu_{ijk}) - d_k.$$

Step C. Substitute (O.1) with $\varepsilon^7 = 0$, so $a_{ij}^7 = g_{ij}^7 + s_j$:

$$g_{ijk}^8 = c + \lambda(g_{ij}^7 + s_j) - d_k + \nu_{ijk}.$$

Step D. Distribute λ and collect:

$$\boxed{g_{ijk}^8 = c + \lambda g_{ij}^7 + \lambda s_j - d_k + \nu_{ijk}.} \quad (\text{O.6})$$

With no measurement error the regressor carries no noise, so $\tilde{g}^7 = \tilde{a}^7$ after the within-transform. Then

$$\hat{\beta}_1 \xrightarrow{p} \frac{\text{Cov}(\tilde{g}^7, \lambda \tilde{g}^7 + \tilde{\nu})}{\text{Var}(\tilde{g}^7)} = \lambda,$$

since $\text{Cov}(\tilde{g}^7, \tilde{\nu}) = 0$ by the classical assumptions. OLS recovers the structural persistence λ without attenuation. The origin fixed effect now targets λs_j , not s_j : persistence scales the difficulty correction.

O.6 Reduced form 4: general model ($\beta_1 = \lambda\rho$)

Reality combines both mechanisms: marks are noisy and the transition has $\lambda < 1$.

Step A. Start from (O.2):

$$g_{ijk}^8 = a_{ijk}^8 - d_k - \varepsilon_{ijk}^8.$$

Step B. Substitute (O.3):

$$g_{ijk}^8 = (c + \lambda a_{ij}^7 + \nu_{ijk}) - d_k - \varepsilon_{ijk}^8.$$

Step C. Substitute (O.1), $a_{ij}^7 = g_{ij}^7 + s_j + \varepsilon_{ij}^7$:

$$g_{ijk}^8 = c + \lambda(g_{ij}^7 + s_j + \varepsilon_{ij}^7) + \nu_{ijk} - d_k - \varepsilon_{ijk}^8.$$

Step D. Distribute λ and collect; name the composite error $\varepsilon_{ijk}^{**} \equiv \lambda\varepsilon_{ij}^7 - \varepsilon_{ijk}^8 + \nu_{ijk}$:

$$\boxed{g_{ijk}^8 = c + \lambda g_{ij}^7 + \lambda s_j - d_k + \varepsilon_{ijk}^{**}} \quad (\text{O.7})$$

Probability-limit derivation. The structural coefficient is λ , but ε^7 appears in the regressor (as $-\varepsilon^7$ inside g^7) and in the error (as $+\lambda\varepsilon^7$ inside ε^{**}). After the within-transform, $\tilde{g}^7 = \tilde{a}^7 - \tilde{\varepsilon}^7$:

$$\hat{\beta}_1 \xrightarrow{p} \lambda + \frac{\text{Cov}(\tilde{g}^7, \tilde{\varepsilon}^{**})}{\text{Var}(\tilde{g}^7)}.$$

Expanding $\text{Cov}(\tilde{g}^7, \tilde{\varepsilon}^{**}) = \text{Cov}(\tilde{a}^7 - \tilde{\varepsilon}^7, \lambda\tilde{\varepsilon}^7 - \tilde{\varepsilon}^8 + \tilde{\nu})$: by the classical assumptions only the $\text{Cov}(-\tilde{\varepsilon}^7, \lambda\tilde{\varepsilon}^7)$ term is non-zero, giving $-\lambda \text{Var}(\tilde{\varepsilon}^7)$. Therefore:

$$\hat{\beta}_1 \xrightarrow{p} \lambda - \frac{\lambda \text{Var}(\tilde{\varepsilon}^7)}{\text{Var}(\tilde{a}^7) + \text{Var}(\tilde{\varepsilon}^7)} = \lambda \left(1 - \frac{\text{Var}(\tilde{\varepsilon}^7)}{\text{Var}(\tilde{a}^7) + \text{Var}(\tilde{\varepsilon}^7)} \right) = \lambda \cdot \underbrace{\frac{\text{Var}(\tilde{a}^7)}{\text{Var}(\tilde{a}^7) + \text{Var}(\tilde{\varepsilon}^7)}}_{\rho}.$$

$$\boxed{\hat{\beta}_1 \xrightarrow{p} \lambda\rho.} \quad (\text{O.8})$$

The estimated slope is persistence times reliability. From $\hat{\beta}_1 \approx 0.60$ alone, λ and ρ are not separately identified: any pair with $\lambda\rho = 0.60$ fits the data equally. However, since both factors lie in $(0, 1]$, each is bounded below by the product: $\lambda \geq 0.60$ and $\rho \geq 0.60$. This one-sided bound requires no additional assumptions beyond the model above.

O.7 The $\hat{\alpha}_j$ continuum indexed by β_1

The four reduced forms imply that $\hat{\alpha}_j$ targets different objects depending on which mechanism dominates. To first order, net of destination effects:

$$\hat{\alpha}_j \approx \underbrace{\overline{g_{ij}^8}}_{\text{cohort anchor performance}} - \beta_1 \underbrace{\overline{g_{ij}^7}}_{\text{cohort raw mark}}. \quad (\text{O.9})$$

The three reference points are:

- **At $\beta_1 = 0$** (no lagged control): $\hat{\alpha}_j = \overline{g_{ij}^8}$, the cohort's mean performance on the anchor, net of destination. This is the calibrated learning level on the Grade 8 scale.

- **At $\beta_1 = 1$ (benchmark):** $\hat{\alpha}_j = s_j$, pure difficulty. The recovery formula $g^7 + \hat{\alpha}_j$ then reconstructs the latent.
- **At $\beta_1 \approx 0.60$ (what the data yield):** $\hat{\alpha}_j$ is a blend: anchor performance net of a 0.6-fraction of the raw Grade 7 mean, a value-added-like quantity in the cross-sectional partialling sense. Of two schools with equal Grade 8 anchor performance, the one with harder assessments (lower raw Grade 7 marks) receives the higher $\hat{\alpha}_j$. This is an algebraic property of the estimator, not a causal production claim. Under equation (O.1) the blend is $\beta_1 s_j + (\lambda - \beta_1) \bar{a}_j^7$, the difficulty term shrunk by the reliability plus a term in the cohort's mean latent Grade 7 learning $\bar{a}_j^7$; that intake term supplies 13–15% of the no-fee contrast, less than the shrinkage hides ([Appendix N](#)).

The strong negative within-cell correlation between $\hat{\alpha}_j$ and raw Grade 7 means (approximately -0.58) is partly mechanical: equation (O.9) subtracts $\beta_1 \bar{g}_j^7$ directly. The residual genuine variation reflects differences in difficulty among schools whose students share similar Grade 8 anchor performance.

Level recovery. Under the general model, the model-consistent fitted level for student i is

$$\text{CalibratedScore}_i = \beta_1 \cdot \text{Gr7Term4}_i + \hat{\alpha}_j,$$

not $\text{Gr7Term4}_i + \hat{\alpha}_j$ (the $\beta_1 = 1$ formula). For rank-based or standardised uses the constant rescaling by β_1 is harmless; for cardinal level recovery it is not.

Exactness caveat. $\hat{\alpha}_j$ from the two-way (origin $\times$ destination) estimator is fixed by the full system of normal equations across the connected mobility network, not literally by $\bar{g}_j^8 - \beta_1 \bar{g}_j^7$ school by school. Equation (O.9) is the correct first-order intuition and the right way to read the dependence on β_1 ; the precise object requires the two-way normal equations. The qualitative continuum in β_1 is exact.

O.8 Technical summary

Model	Transition	Marks	Plim of $\hat{\beta}_1$	Origin FE targets	$\beta_1 < 1$?
O.3 benchmark	level-preserving ($\lambda = 1$)	exact	1	s_j (difficulty)	no
O.4 pure attenuation	level-preserving ($\lambda = 1$)	noisy	ρ	s_j , attenuated	yes, measurement artefact
O.5 pure fade	partial ($\lambda < 1$)	exact	λ	λs_j	yes, genuine dynamics
O.6 general	partial ($\lambda < 1$)	noisy	$\lambda \rho$	$\approx \lambda s_j$, blended	yes, both

The headline estimated slope $\hat{\beta}_1 \approx 0.60$ sits in the last row. Decomposing λ from ρ requires extra information; [Appendix N](#) uses the English mark to bound them at $\lambda \geq 0.75$ and $\rho \leq 0.70$ in 2018, and $\lambda \geq 0.74$ and $\rho \leq 0.77$ in 2022, with attenuation the larger factor in 2018 and about equal to fade in 2022. The structural equations, the identity (O.8), and the $\hat{\alpha}_j$ continuum are summarised for the main text in §3.1 of the main text.

O.9 The complete set of identifying assumptions

Sections O.1–O.2 fixed the structural primitives and §O.4–O.6 the classical errors-in-variables conditions that govern the slope. For completeness we now collect *every* assumption the calibrated estimator requires, grouped by the layer of the design it disciplines. The conditions named in the main text are economic

summaries of this set; the remainder are either formal regularity conditions or design properties established elsewhere in the paper. None is assumed in isolation: each is matched to a design feature, a robustness check, or a falsification test.

Axiom	Statement	Role, and where disciplined
M1	Difficulty is additive with unit loading and constant within school: $a_{ij}^7 = g_{ij}^7 + s_j + \varepsilon_{ij}^7$, with no school-specific scale stretching ($\theta_j \equiv 1$) and s_j independent of a student's position in the mark distribution (O.1).	Gives a single slope and a scalar α_j per school. A school-specific θ_j would appear as an origin-specific slope (§O.10). Slopes do differ by origin, through within-school reliability rather than θ_j , and the order of schools at the average mark barely moves (Appendix S.5).
M2	Destination offset d_k is additive, common to all students in cell k , and independent of origin j (O.2).	Makes γ_k a genuine destination offset. This is the main text's <i>common within-grade anchor</i> ; warranted by one assessment per cell under common conditions.
M3	Classical measurement error: $\varepsilon^7, \varepsilon^8$ are mean-zero, mutually uncorrelated, and uncorrelated with latent learning and the growth shock (§O.1).	Fixes the slope's attenuation to the scalar ρ (O.5, O.8). Relaxations are examined in Appendix N .
D1	Homogeneous linear persistence: a single $\lambda \in (0, 1]$, constant across students, origins and destinations (O.3).	<i>Generates</i> additive separability (§O.10) and the single slope. A destination-specific λ_k breaks separability; the slope does differ across destinations, but fitting it barely reorders feeders (§O.10).
D2	Growth-shock exogeneity: $E[\nu_{ijk} \mid a_{ij}^7, \varepsilon^7, \varepsilon^8] = 0$.	With M3, delivers the probability-limit algebra of §O.6.
X1	Exogenous mobility: conditional on (g_{ij}^7, j, k) , the destination a student enters is independent of the composite residual $\varepsilon_{ijk}^{**} = \lambda \varepsilon^7 - \varepsilon^8 + \nu$, equivalently no sorting on unobserved growth.	The AKM identifying condition; keeps $\hat{\alpha}_j$ free of selection. Sorting on g^7 is permitted and absorbed; tested in the main text and Appendix F , Appendix H .
X2	Connectedness: within each province-cohort cell the bipartite origin $\times$ destination graph has a giant connected component; relative α_j are identified only within it.	Chains the within-grade comparisons into one scale. Coverage and the λ_2 density argument: main text and Appendix L .
S1	Non-selective linkage: being observed in both grades within province is independent of $\{\nu_{ijk}, s_j\}$ given the fixed effects.	Makes $\hat{\alpha}_j$ valid for the full population of origin j , not only its linked sub-sample. Attrition robustness: Appendix A .

Axiom	Statement	Role, and where disciplined
R1	Large feeders, dense network: each origin links many students to many destinations, so the origin-mean measurement error $\bar{\varepsilon}_j^7 \rightarrow 0$ and limited-mobility bias is negligible.	Finite-sample noise in $\hat{\alpha}_j$ averages out, and ranks survive attenuation under near-uniform reliability. Signal-to-noise ≈ 7 in the main text; λ_2 benchmark in Appendix L .

In these terms the main text’s identifying conditions are exogenous mobility (X1), within a connected set (X2); additivity, which §O.10 derives from M1, M2 and D1; the exclusion restriction, which no axiom above states; and, among the conditions specific to school-set marks, the common within-grade anchor (M2) and class marking unrelated to the feeder. M3 and D2 are the classical conditions of §O.1 that the plim algebra consumes; S1 and R1 are regularity conditions. Sections O.3–O.6 derived the estimating equation taking M1–M3 and D1–D2 as given; the assignment, sample and regularity axioms (X1–X2, S1, R1) are what carry that algebra from the population to the estimated $\hat{\alpha}_j$.

O.10 Additive separability as a consequence, and what breaks it

The main text lists additive separability of α_j and γ_k among its identifying conditions. Under the primitives of §O.9 it is not an independent assumption but a *consequence* of three of them.

Proposition (additive separability). Under M1, M2 and D1, the Grade 8 mark decomposes as

$$g_{ijk}^8 = c + \lambda g_{ij}^7 + \underbrace{\lambda s_j}_{\alpha_j} - \underbrace{d_k}_{\gamma_k} + \varepsilon_{ijk}^{**},$$

in which the origin enters only through $\alpha_j = \lambda s_j$, the destination only through $\gamma_k = -d_k$, and there is no origin $\times$ destination interaction term.

Proof. This is the general reduced form (O.7). The origin contributes λs_j : by D1 the factor λ carries no destination index, and by M1 the shift s_j carries no destination index, so the origin term is constant across k . The destination contributes $-d_k$, which by M2 carries no origin index and is constant across j . Hence the systematic part is the sum of an origin-only and a destination-only term, with no α_{jk} interaction. ■

The attenuation of the slope to $\lambda\rho$ (O.8) does not disturb separability: ε^7 being common to the regressor and the error rescales the *common* slope but leaves α_j and γ_k additive, and the within-transform annihilates both fixed effects regardless (§O.4). The estimator therefore needs additivity only *in mean*: idiosyncratic origin-by-destination deviations that average to zero across a school’s many destinations are harmless; only a *systematic* interaction biases $\hat{\alpha}_j$. In the data the idiosyncratic deviations are small: net of estimation noise their SD is 0.079 within-province SD in 2018 and 0.078 in 2022 on the links that two or more students take, and they do not line up with how hard either school’s assessments are ([Appendix S.5](#)).

Corollary (what breaks it). A systematic interaction arises exactly when D1 or M1 fails.

- (i) *Heterogeneous persistence.* If $\lambda = \lambda_k$ varies by destination, Step C of §O.6 instead yields

$$g_{ijk}^8 = c + \lambda_k g_{ij}^7 + \lambda_k s_j - d_k + \varepsilon_{ijk}^{**},$$

so the origin term $\lambda_k s_j$ is an origin $\times$ destination interaction and the slope is destination-specific; the two-way additive specification is then misspecified.

- (ii) *Scale stretching.* If difficulty loads non-unitarily, $a_{ij}^7 = \theta_j g_{ij}^7 + s_j + \varepsilon_{ij}^7$ with $\theta_j \neq 1$, the slope becomes origin-specific ($\lambda\theta_j$) and $\hat{\alpha}_j$ depends on the within-cell shape of a school’s mark distribution rather than on a single location shift.

Both departures are restrictions on the *primitives*, not on the estimator. The short window does not rule out case (i): the slope of the Grade 8 mark on the Grade 7 mark differs across secondary schools, with a pupil-weighted standard deviation of 0.18 in 2018 and 0.14 in 2022 around a mean near 0.58, and a school’s slope in one cohort correlates 0.43 with its slope in the other. A secondary that stretches its own marks produces the same pattern, so the data cannot say whether persistence or difficulty differs. Either way the wedges barely move: fitting case (i) in full, with each secondary’s slope scaling the feeder’s offset as well as the Grade 7 mark, gives rank correlations with the baseline of 0.981 in 2018 and 0.994 in 2022, and a spread of the wedges 6 and 3 per cent narrower. Case (i) would make a feeder’s effect differ by destination, yet no cell of the mean residuals by wedge decile and destination decile departs from zero by more than 0.039 within-province SD in 2018 or 0.037 in 2022, although the pair interactions are jointly significant in 2018 ([Appendix S.5](#)). The main text’s mean residuals by the student’s Grade 7 decile and the wedge decile show a feeder’s wedge varying with the mark. A slope per feeder takes up most of that variation and leaves the order of feeders at the average mark nearly unchanged. Case (ii) does not produce the pattern: a school-specific θ_j would act through the ratio of a feeder’s Grade 8 spread to its Grade 7 spread, and that ratio accounts for little of the slopes’ rise with the wedge in 2018 and runs against it in 2022. The steeper slopes of feeders with harder assessments come instead from a closer correlation between the two marks, the signature of a higher within-school reliability ρ_j rather than a different scale (§3.3 of the main text; [Appendix S.5](#)).

Appendix P: Wedge-Supply Exhibits

This appendix reproduces the wedge-supply exhibits that support §4.5 of the main text. All exhibits are re-formats of existing wedge-supply outputs; no new estimation is performed here.

Variance anatomy of the school-level mark–learning gap

Table 48: Variance decomposition of the school-level mark–learning gap (full table).

Component	GT (unwtd)	GT (wtd)	6P (unwtd)	6P (wtd)
(i) Estimation noise (mean SE^2)	2.99	2.14	6.17	3.85
(ii) Province + cohort FE	0.00	0.00	0.00	0.35
(iii) District / province	10.55	10.77	4.10	5.24
(iv) Quintile / district	16.84	18.43	6.56	9.68
(v) Persistent school component	30.05	33.31	28.54	31.20
(vi) Transitory residual	39.58	35.35	54.63	49.67
Total Var(wedge), raw units	0.22	0.20	0.21	0.19
Within-district share (%)	89.45	89.23	95.90	94.41
Within-quintile share (%)	72.61	70.80	89.34	84.73
Persistent/within-dist signal share (%)	34.75	38.25	31.81	34.46

Sample: GT pooled (2,903 school-cohorts; $n = 1,214$ matched for cross-cohort covariance); 6P = six-province panel (~22,869 school-cohorts). *Method:* Sequential OLS FE decomposition; persistent component = $Cov(residual\ 2018, residual\ 2022)$ on matched schools; noise = $mean(SE^2)$ [Kane–Staiger].

Covariate atlas, forest plot

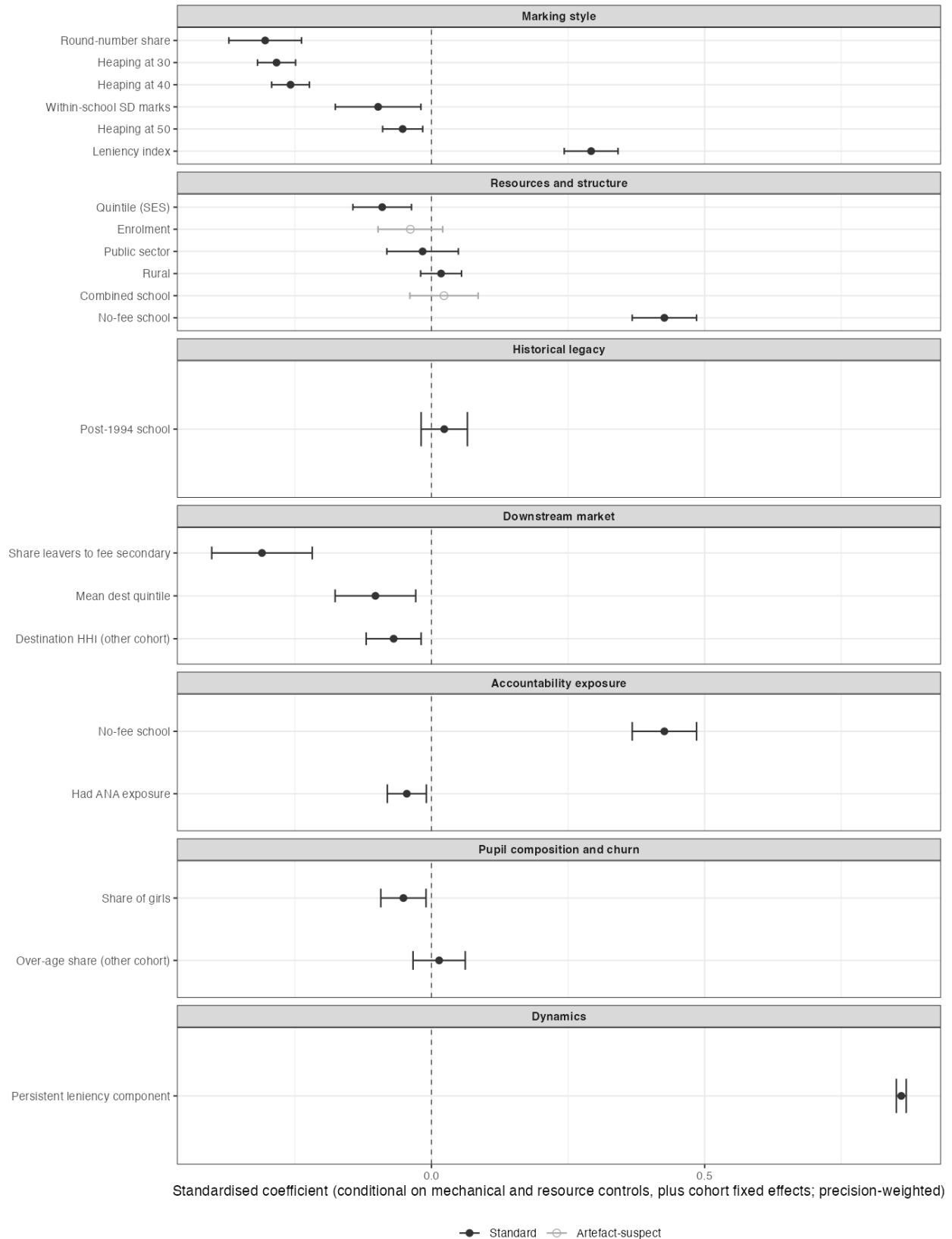


Figure 26: Forest plot of all covariate associations with the school-level mark–learning gap (GT; precision-weighted; each conditional on the mechanical and resource families, the school’s own poverty quintile among them). Covariates whose Romano–Wolf p-value, adjusted within their family, is below 0.05 are marked; among resources and structure, the no-fee indicator and the school’s own poverty quintile both survive the stepdown. Artefact-suspect covariates (is_combined, enrolment) are shown in grey. In the image, “leniency” is the wedge with its sign reversed, so a positive coefficient is easier assessments, and “Leniency index” is the Term 4 lift (term4_lift). Sample: GT pooled (up to 2,753 school-cohorts with an own poverty quintile; 1,000 school-cohorts with no own poverty quintile).

Covariate atlas, variable dictionary

Table 49: Covariate atlas — variable dictionary. Every regressor plotted in Figure 26 and listed in Table 50, followed by the mechanical controls held fixed in each regression.

Variable	Family	Definition
<code>term4_lift</code>	Marking style	School mean of the Grade-7 Term-4 mark minus the school mean of the average of Terms 1 to 3, in raw mark points. Positive: terminal marks lifted above the school's own earlier-term standard.
<code>heap_30</code>	Marking style	Share of the school's Grade-7 Term-4 marks equal to exactly 30 (a progression cut-point).
<code>heap_40</code>	Marking style	Share of Grade-7 Term-4 marks equal to exactly 40.
<code>heap_50</code>	Marking style	Share of Grade-7 Term-4 marks equal to exactly 50.
<code>round_number_share</code>	Marking style	Share of Grade-7 Term-4 marks divisible by 5.
<code>within_school_sd</code>	Marking style	Standard deviation of Grade-7 Term-4 marks within the school.
<code>no_fee</code>	Resources and structure	Indicator: no-fee (fee-exempt) school.
<code>quintile</code>	Resources and structure	National poverty quintile of the school itself (1 = poorest to 5 = least poor), from the masterlist. The sending school's own socioeconomic standing; entered alongside <code>no_fee</code> so every other covariate is conditioned on own SES.
<code>is_public</code>	Resources and structure	Indicator: public rather than independent school.
<code>is_combined</code>	Resources and structure	Indicator: combined-phase school (primary and secondary grades on one campus). Artefact-suspect.
<code>enrolment</code>	Resources and structure	Total student enrolment. Artefact-suspect.
<code>is_rural</code>	Resources and structure	Indicator: rural location.
<code>is_new_school</code>	Historical legacy	Indicator: founding date after 1994 (crude proxy for the apartheid-era department that ran the school).
<code>dest_hhi_other_cohort</code>	Downstream market	Herfindahl concentration of the secondaries the school's leavers enter, from the other cohort's flows.
<code>mean_dest_quintile</code>	Downstream market	Mean poverty quintile of the destination secondaries the leavers enter.
<code>share_leavers_fee_secondaries</code>	Downstream market	Share of leavers enrolling in fee-charging secondaries.
<code>had_ana</code>	Accountability exposure	Indicator: school participated in the Annual National Assessments. (<code>no_fee</code> also enters this family as an accountability proxy.)
<code>overage_share_other_cohort</code>	Student composition and churn	Share of students above the modal age for their grade, from the other cohort.
<code>gender_ratio_f</code>	Student composition and churn	Share of girls among the school's linked Grade 7 Mathematics students in the same cohort.

(continued)

Variable	Family	Definition
wedge_persistent	Dynamics	Cross-cohort average of the school's calibrated wedge (its persistent component). A transform of the outcome; excluded from the Shapley decomposition (Table W.S5) as circular.
wedge_se	Mechanical control	Standard error of the school-cohort wedge estimate (its sampling precision). Also the precision weight, $1/\text{wedge_se}^2$.
n_students	Mechanical control	Students in the school-cohort.
n_gr8_destinations	Mechanical control	Number of distinct destination secondaries.
dest_hhi	Mechanical control	Herfindahl concentration of destinations (own cohort).
reliability_lambda	Mechanical control	Signal reliability: signal variance / (signal variance + mean squared standard error).

Specification. The plotted rows (Figure 26) and Table 50 regress the sign-reverse of the standardised per-cohort wedge (*wedge_std*) on each standardised covariate; the joint R^2 and the Shapley shares of Table 51 use the standardised persistent component (*wedge_persistent*). Every regression weights school-cohorts by $1/\text{wedge_se}^2$, absorbs cohort fixed effects, holds the mechanical and resource families fixed, and clusters standard errors at the school. A covariate that is itself a control is dropped from the control set on its own row, so the no-fee coefficient is its own-family resource association rather than a stepdown survivor.

Covariate atlas, full coefficient table

Table 50: Full covariate atlas: associations with the school-level mark-learning gap across four specifications and two weighting schemes. Romano-Wolf adjusted p-values from 1,000 school-cluster resamples.

Covariate	Family	Coefficient	SE	p	RW p	Artefact
wedge_persistent	Dynamics	0.857	0.005	< 0.001	0.000	No
heap_30	Marking style	-0.265	0.018	< 0.001	0.000	No
heap_40	Marking style	-0.222	0.018	< 0.001	0.000	No
heap_50	Marking style	-0.021	0.022	0.344	0.005	No
term4_lift	Marking style	0.332	0.030	< 0.001	0.000	No
round_number_share	Marking style	-0.349	0.038	< 0.001	0.000	No
within_school_sd	Marking style	-0.282	0.040	< 0.001	0.005	No
enrolment	Resources and structure	0.075	0.024	0.002	0.566	Yes
is_combined	Resources and structure	-0.110	0.026	< 0.001	0.691	Yes
is_public	Resources and structure	0.240	0.029	< 0.001	0.691	No
is_rural	Resources and structure	0.076	0.024	0.002	0.691	No
no_fee	Resources and structure	0.490	0.022	< 0.001	0.000	No
quintile	Resources and structure	-0.348	0.022	< 0.001	0.008	No
is_new_school	Historical legacy	0.028	0.024	0.238	0.254	No

(continued)

Covariate	Family	Coefficient	SE	p	RW p	Artefact
dest_hhi_other_cohort	Downstream market	-0.117	0.026	< 0.001	0.007	No
mean_dest_quintile	Downstream market	-0.366	0.024	< 0.001	0.007	No
share_leavers_fee_secondaries	Downstream market	-0.502	0.022	< 0.001	0.000	No
had_ana	Accountability exposure	0.135	0.047	0.004	0.024	No
no_fee	Accountability exposure	0.490	0.022	< 0.001	0.000	No
gender_ratio_f	Pupil composition and churn	-0.086	0.023	< 0.001	0.087	No
overage_share_other_cohort	Pupil composition and churn	0.176	0.029	< 0.001	0.558	No

Every column is precision-weighted; Coefficient, SE and p are from the specification without cohort fixed effects and RW p is the Romano–Wolf adjusted p-value. Artefact flags the two covariates the variable dictionary (Table 49) marks artefact-suspect. Sample: GT pooled cohorts (2,903 school-cohorts; 1,280 GT schools with both cohorts matched). Precision-weighted = $1/\text{wedge_se}^2$. All continuous predictors standardised (z-scores). F1 controls: wedge_se, n_students, n_gr8_destinations, dest_hhi, reliability_lambda. Romano–Wolf: 1,000 reps, school-cluster resample, per-family stepdown. Associations only; no causal interpretation.

Shapley variance decomposition

Table 51: Shapley decomposition of persistent-gap variance by covariate family (GT matched panel, n = 1,280 schools).

Family	Shapley share	Joint R ² (%)	Variables retained
Downstream market	0.1422	38.3	destination concentration; mean destination quintile; share of leavers to fee-charging secondaries
Resources and structure	0.1070	28.8	no-fee status; own poverty quintile; public school; combined school; enrolment; rural location
Accountability exposure	0.1056	28.4	participated in the ANA; no-fee status
Pupil composition and churn	0.0098	2.6	over-age share; share of girls
Historical legacy	0.0067	1.8	founded after 1994

Sample: GT Mathematics matched-panel schools (both cohorts; n = 1,280). Outcome: wedge_persistent (standardised cross-cohort average). Within-family screening retains every covariate correlated with the outcome above 0.05 in absolute value, which in practice retains all of them. The families are not disjoint: the no-fee indicator enters both resources and structure and accountability exposure, so those two shares each carry part of the same variable. Shapley exact enumeration (2^K , K = 5 families F3–F8). Joint R² = 0.371. Shares sum to the joint R²; no causal interpretation.

Hypothesis tests H1 and H2

The downstream market is the set of secondary schools a primary school’s leavers enter. Its socioeconomic composition tracks the wedge more closely than anything else about it: the harder a school’s assessments,

the richer the schools its students enter, and the less likely it is to send them to a no-fee secondary school. Both associations hold the school's own poverty quintile and fee status fixed, so destination wealth tracks difficulty over and above the school's own standing. Concentration of the market carries far less. Where a school's leavers pool into few destinations the school's assessments are harder: each one-standard-deviation rise in destination concentration, measured as the other-cohort Herfindahl index of the schools a school's leavers attend, is associated with a wedge 0.033 SD higher, 0.07 of the wedge's spread across schools (Gauteng precision-weighted; SE 0.012, $p = 0.0059$; Table 52), about a sixth of the no-fee gradient.

The correlation admits four readings, which I do not separate. The first is precision mechanics. A more concentrated destination market goes with a larger downstream network, and network size sharpens the precision of $\hat{\alpha}_j$, so part of the correlation may be an artefact of estimation. The second is reverse causation. The leavers of a school with harder assessments are admitted more readily by particular secondary schools, so they pool there. The third is a common cause driving both. The fourth is a repeated game: a primary school facing a concentrated secondary market sets harder assessments because that market can hold it to account in the next round.

Table 52: Hypothesis test results: H1 (internalisation) and H2 (reputational discipline) across four sample specifications.

Hypothesis	Sample	Coef	SE	p-value	95% CI	Adjudication
H1: combined (main effect)	GT weighted	-0.101	0.129	0.435	[-0.354, 0.152]	underpowered
H1: combined (main effect)	GT unweighted	-0.067	0.134	0.615	[-0.329, 0.195]	
H1: combined (main effect)	6P weighted	-0.130	0.020	< 0.001	[-0.170, -0.090]	
H1: combined (main effect)	6P unweighted	-0.078	0.015	< 0.001	[-0.108, -0.048]	
H1: combined $\times$ own retention (interaction)	GT weighted	0.060	0.042	0.150	[-0.022, 0.141]	supported
H1: combined $\times$ own retention (interaction)	GT unweighted	0.031	0.045	0.491	[-0.057, 0.120]	
H1: combined $\times$ own retention (interaction)	6P weighted	0.007	0.009	0.424	[-0.010, 0.025]	
H1: combined $\times$ own retention (interaction)	6P unweighted	-0.029	0.007	< 0.001	[-0.043, -0.016]	
H2: destination concentration	GT weighted	-0.033	0.012	0.006	[-0.056, -0.009]	
H2: destination concentration	GT unweighted	-0.032	0.011	0.004	[-0.054, -0.010]	
H2: destination concentration	6P weighted	-0.022	0.005	< 0.001	[-0.032, -0.012]	
H2: destination concentration	6P unweighted	-0.014	0.005	0.003	[-0.024, -0.005]	

Sample: GT = Gauteng pooled school-cohorts with an own poverty quintile and an other-cohort destination market (precision-weighted and unweighted); 6P = the six-province panel on the same restrictions (province $\times$ cohort FE). H2 controls for the school's own poverty quintile alongside no-fee status; H1 keeps its original specification. All continuous predictors standardised (z-scores). Outcome = the sign-reverse of the wedge, in SD of Grade 8 marks (not standardised across schools). H1 is artefact-suspect (pct_drop_wtd = 78.7% under F1 controls); treat H1 results as underpowered. H2 is not artefact-suspect (pct_drop_wtd = 50.9%, below the 75% threshold). Associations only; no causal interpretation.

Spatial clustering

The wedge clusters in space. Neighbouring schools grade alike, with a Moran's I of about 0.197 at a 5 km radius (Gauteng, 2018), which is consistent with assessment norms diffusing between neighbours or with spatially correlated unobservables.

Table 53: Spatial autocorrelation and neighbour-mean associations in the school-level mark-learning gap.

Specification	Coef / Stat	SE	p-value
Moran's I (GT, 5 km, 2018)	0.1971	—	—
Moran's I (GT, 10 km, 2018)	0.1615	—	—
Moran's I (GT, 5 km, 2022)	0.2250	—	—
Moran's I (GT, 10 km, 2022)	0.1908	—	—
Neighbour-mean (GT, wtd, cond. F1+F3+district FE)	0.3222	0.0367	< 0.001
Neighbour-mean (GT, uwtd, cond. F1+F3+district FE)	0.3243	0.0377	< 0.001
Neighbour-mean (6P, uwtd, cond. F1+F3+prov×coh+district FE)	0.3470	0.0150	< 0.001
Dest-mean (GT, wtd, cond. F1+F3+district FE)	0.2600	0.0298	< 0.001
Dest-mean (GT, uwtd, cond. F1+F3+district FE)	0.2563	0.0285	< 0.001

Every row is consistent with either the diffusion of assessment norms or spatially correlated unobservables; the design does not separate them. Sample: GT pooled cohorts headline ($n = 2,700$ school-cohorts with GIS; GIS coverage 93%); six-province robustness ($n = 21,889$ with GIS). Moran's I computed with binary spatial weights at stated band. Neighbour-mean: 5 nearest neighbours (Haversine, leave-self-out). Dest-mean: feeders sharing modal Gr8 destination (leave-self-out). Controls: F1 + F3 covariates + district FE. Cluster SE at school.

Spatial distribution of the gap (GT map)

Dynamics, gap change 2018 to 2022

Between cohorts the wedge reverts towards the cohort mean. The cross-cohort convergence coefficient is -0.513: a school far from the mean in 2018 has travelled about halfway back by 2022. Measurement error alone would produce some reversion, but the measure's reliability (§4.2 of the main text) puts that mechanical floor near zero. The coefficient stands as an upper bound on the strength of true convergence rather than an artefact of noise reverting on itself.

Table 54: Dynamics of the school-level mark-learning gap: OLS regression of gap change (2022 – 2018) on baseline covariates.

Term	GT (wtd)	GT (unwtd)	6P (wtd)	6P (unwtd)
d_enrolment	0.0001	-0.0000	-0.0001	-0.0002*
baseline_no_fee	0.3495***	0.3620***	0.2867***	0.3043***
baseline_combined	-0.1109.	-0.1363*	-0.0412	-0.0423
baseline_wedge	-0.5134***	-0.5388***	-0.5703***	-0.6299***

Sample: GT matched-panel schools ($n = 1,270$ of the 1,280-school matched panel, $\times 2$ cohorts); six-province matched-panel ($n = 9,892$ schools). LHS: the wedge's change, 2022 minus 2018, sign reversed. Weights: 1/baseline_wedge_se² (precision-weighted). Cluster-robust SEs at school. Province FE in all specifications. baseline_wedge = convergence term: negative



Figure 27: Geographic distribution of the school-level mark–learning gap across GT schools. Colour encodes the persistent wedge, sign reversed (precision-weighted cross-cohort average); darker = larger gap. In the image, “leniency” is the wedge with its sign reversed: positive is easier assessments. Spatial clustering (Moran’s $I \approx 0.20$ at 5 km in 2018) is consistent with either norm diffusion or spatially correlated unobservables; not separable here. *Sample*: GT matched-panel schools ($n = 1,280$; both cohorts; GIS-matched).

coefficient indicates mean-reversion; this is partly mechanical noise-reversion. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.10$.

Gap dynamics, kernel density 2018 vs 2022

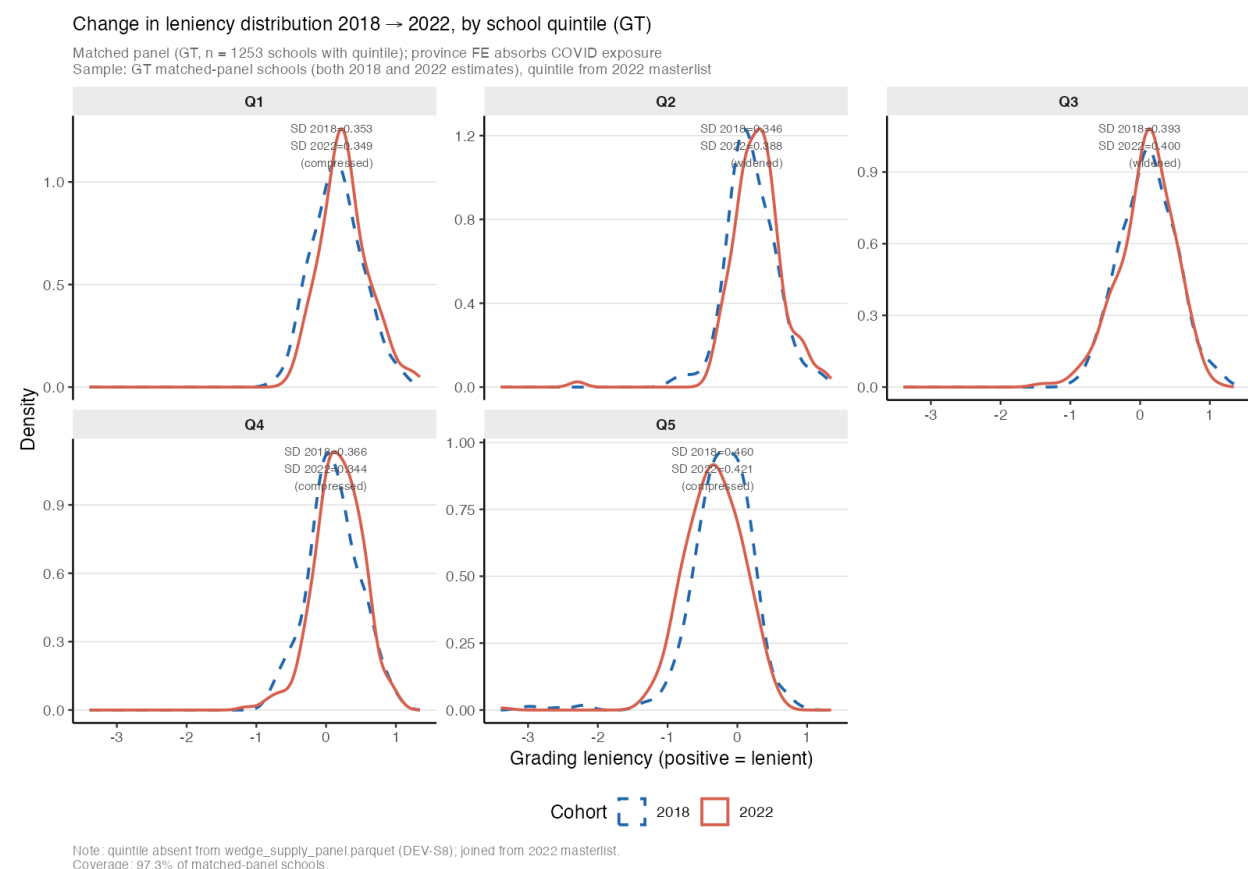


Figure 28: Kernel densities of the school-level mark–learning gap in 2018 and 2022, by poverty quintile (GT matched-panel schools with a quintile). Each cohort’s wedge is in that cohort’s own SD, so the two spreads are not directly comparable. In the image, “leniency” is the wedge with its sign reversed: positive is easier assessments. *Sample*: GT Mathematics matched-panel schools with a poverty quintile (precision-weighted).

Natural-experiment inventory (L1–L11)

The full 11-lever inventory assessing candidate identification strategies for the causal analogue of §4.5 of the main text is reproduced below. Levers are rated on a 1–3 feasibility scale (1 = high; 3 = low).

Lever	Description	Feasibility
L1	Gauteng online-admissions rollout (staggered DiD)	2
L2	Provincial moderation-regime differences	3
L3	Jika iMfundo district rollout (KZN)	2
L4	Fee-status and quintile reclassification events	2
L5	ANA suspension 2016–2018	2
L6	School merger / closure events	3

Lever	Description	Feasibility
L7	Teacher union strike disruptions	3
L8	School principal replacement events	2
L9	Umalusi moderation shock (province $\times$ year)	2
L10	Introduction of fee-charging secondaries in feeder zones	2
L11	Satellite school de-linking events	2

Feasibility scale: 1 = high (actionable within 6 months with available/public data; clean design); 2 = medium (actionable with RA effort or a data request; design has solvable complications); 3 = low (requires major acquisition, regulatory access, or has fundamental identification threats). Two of these levers cover the two largest associations in §4.5 of the main text, the no-fee gradient and the market-structure coefficient.

Fee and quintile reclassifications (L4) move a school’s fee status while its intake stays fixed, and the data are already in hand. Comparing the wedge before and after a reclassification would test the no-fee gradient against a change in fee status that leaves the student composition alone.

The entry of a fee-charging secondary school into a feeder zone (L10) changes the concentration of the destination market that an existing school’s leavers face. Tracking the wedge around such entries would test the market-structure association against a shift in downstream concentration that a school does not choose. Both designs turn on a change that the affected schools do not select into, which is what the correlations in §4.5 of the main text cannot supply.

Appendix Q: Comparability frameworks

This appendix delimits the relationship between the present exercise and the psychometric literature on test linking and equating. Its purpose is taxonomic, not methodological: it places the calibration in the correct conceptual bin and identifies what claims the bin warrants and what it does not.

The Mislevy–Linn taxonomy

The canonical treatment of test linking (Kolen and Brennan 2014) distinguishes four linking goals along a continuum from weakest to strongest warrant:

1. **Moderation.** Adjusting marks from different assessments to a common scale when the assessments are not built to common specifications. Interchangeability of scores is not warranted.
2. **Calibration.** Placing scores from assessments measuring related but non-identical constructs on a common scale. Scores are comparable, not interchangeable.
3. **Projection.** Predicting performance on one assessment from scores on another, where the assessments measure related but distinct constructs.
4. **Equating.** The strongest form; requires assessments built to common specifications so that scores are interchangeable in the strict sense.

By this taxonomy, the present exercise is **statistical moderation** in the Mislevy–Linn sense, following Kolen and Brennan (2014, sec. 10.1.2): the Grade 7 school-based assessments are not built to common cross-school specifications; accordingly equating in the strict sense is precluded.

The Holland–Dorans goal-based framework

Holland and Dorans (2006) organise linking goals into three categories: *predicting*, *scale aligning*, and *equating*. The present exercise is closest to scale aligning, placing each school’s marks on the scale defined by the Grade 8 within-grade anchor, rather than predicting or equating.

Lord’s equity and what the calibration does not claim

Kolen and Brennan (2014, sec. 1.3) formalise Lord’s equity property: true equating requires that the distribution of scores for any subgroup of examinees be the same regardless of which form they sat. Lord’s

equity is known to be unattainable except for essentially identical test forms, a condition that no setting linking heterogeneous school-based assessments can satisfy. The calibrated school effect $\hat{\alpha}_j$ is a school-level location parameter, not a student-level interchangeable score; no claim to Lord’s equity is made or implied.

The Masters (1988) anchor-validity criterion

Masters (1988) establishes the validity criterion that governs every application of statistical moderation: the anchor must measure the same variable as the assessments being rescaled. Masters’ empirical illustration of the failure mode (the 1984 ACT attempt to rescale within-school English assessments against a verbal–quantitative aptitude test, which produced sex bias because the quantitative section did not measure English-language attainment) shows that anchor selection can distort results when the anchoring instrument conflates distinct constructs. The subgroup-invariance check reported in [Appendix H.3](#) is the operational diagnostic that the Masters condition holds in the present setting: the estimated translation constant is stable across school-type, quintile, and province subgroups.

Placement: statistical moderation through a within-grade anchor

Taken together, the taxonomy places the calibration as follows:

- It is **not equating** (Grade 7 assessments are not built to common cross-school specifications; Lord’s equity is not available; the inter-assessment correlation is school-level, not item-level).
- It is **not projection** (the exercise does not predict Grade 8 performance from Grade 7 marks; it uses Grade 8 outcomes to construct a cross-school common scale for Grade 7 school-of-origin effects).
- It is **statistical moderation** (Mislevy–Linn) / **scale aligning** (Holland–Dorans): the Grade 8 within-grade assessment acts as the anchor, and the school fixed effect $\hat{\alpha}_j$ is the moderation constant. A student’s calibrated score is her Grade 7 mark times the Grade 7 slope $\hat{\beta}_1$ plus her school’s $\hat{\alpha}_j$, and that sum places each school’s marks on the anchor’s scale.

The two distinctions that matter for interpretation are: (a) linking adjusts for differences in assessment difficulty and examinee ability across forms, warranting cross-school comparability but not interchangeability; and (b) the estimand is a scale parameter ($\hat{\alpha}_j$), not a causal quantity, and the wedge is a school-level location parameter; the calibrated score built from it compares students across schools but is not an equated test score. Both distinctions are stated in §3 of the main text and carried through the rest of the paper.

Appendix R: Checks on the exit-school effects

This appendix lists the checks behind §6.2 of the main text.

Table 55: **Effects pass the checks on signal and most on sorting; one cohort limits the rest.**

Check	Statistic	Mathematics	Mathematical Literacy	Threshold	Result	Liu et al. (2025)
Signal	split-half reliability of θ	0.855	0.841	≥ 0.60	passes	year-to- year correla- tion instead
Signal	signal-to- noise ratio	2.450	2.403	≥ 2.0	passes	not reported
Specifi- cation	lowest agreement across choices that leave the intake unchanged	0.972	0.973	≥ 0.80	passes	outcome defini- tions varied

(continued)

Check	Statistic	Mathematics	Mathematical Literacy	Threshold	Result	Liu et al. (2025)
Attrition	agreement with the median attrition scenario	0.861	0.801	≥ 0.80	passes	not reported
Subject choice	agreement with subject-choice reweighting	0.995	0.996	≥ 0.90	passes	not reported
Placebo, raw mark	dispersion ratio, raw Grade 7 Term 1 mark	1.061	1.282	≤ 1.00	fails	not reported
Placebo, calibrated mark	the same, calibrated, same students	0.364	0.434	≤ 0.50	passes	not applicable
Placebo, held out	largest of three placebos the intake has not seen	0.971	1.237	≤ 0.50	fails	not reported
Convergence	movement of θ , last two steps of the intake ladder	0.075; 0.014	0.064; 0.034	none	reported	not reported
Within-cohort forecast	corrected coefficient, half on half	1.030 [1.011, 1.146]	1.051 [0.996, 1.243]	0.85 to 1.15	reported	leave-one-year-out prediction instead
Non-choosers	calibrated placebo among students who did not choose their school	0.675	0.688	≤ 0.50	fails	not reported
Measured sorting	projection share of θ on the calibrated placebo; worst case	0.018; 0.186	0.028; 0.296	none	reported	composition correlation instead

(continued)

Check	Statistic	Mathematics	Mathematical Literacy	Threshold	Result	Liu et al. (2025)
Adjusted bands	schools in the same band: measured-sorting adjustment; proportional-selection bound	0.951; 0.780	0.951; 0.777	≥ 0.80 , second only	second fails	composition correlation instead
Cross-cohort	school-level agreement of an earlier cohort's residual with θ	0.518	0.536	0.60	below	several cohorts, same measure
Precision	schools whose above/middle/below band is unstable across draws	0.402	0.386	≤ 0.20	fails	not reported
Link	link rate; quintile-5 over quintile-1 feeders	0.637; 0.239/0.756	0.637; 0.239/0.756	≥ 0.70	fails	not applicable

Notes: Each row is one check on the exit-school effects θ_k . Mathematics and Mathematical Literacy are separate estimations. Thresholds are the levels each check was set to meet; ‘none’ marks a check with no threshold, and ‘below’ names the level a student-level cross-cohort test would need. The forecast coefficient uses the correction that does not share the estimation sample of θ_k ; a simpler correction sits at one by construction. The calibrated placebo test and its 0.50 ceiling were specified after the raw placebo test had failed. The last column records whether Liu et al. (2025), the closest design, report the same check; ‘instead’ names the check they report in its place, and ‘not applicable’ marks a check their external intake makes unnecessary. The row-by-row term search behind it is `tabV2_liu_check.csv`. The agreement with the median attrition scenario for Mathematical Literacy sits within the margin of its threshold. The missing-value glyph for this document is an em dash. None arises here. A placebo’s dispersion ratio is the spread of its estimated effects across secondary schools divided by the spread of the exit-school effects. Sample: six provinces (Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga and North West) and the cohort of Grade 7 2020, Grade 8 2021 and NSC 2025: 394,353 linked students at secondary schools with at least twenty linked students holding a calibrated Grade 7 intake; 1,135 secondary schools with at least twenty Mathematics students and 1,486 with at least twenty Mathematical Literacy students; and, for the cross-cohort row, 942 secondary schools for Mathematics and 1,272 for Mathematical Literacy of the 2018 to 2023 cohort.

Appendix S: AKM diagnostics

This appendix reports the two-way fixed-effects diagnostics behind Table 4 of the main text and §3.3 of the main text.

S.0: Where each threat is examined

This appendix runs the diagnostics of the two-way fixed-effects (AKM) literature on the wedge (Kline 2024, 140–45; Bonhomme et al. 2026, 16–17). Table 56 lists the standard checks, most of which the sections below detail, then the threats specific to school-set marks, led by the one those checks cannot detect: whether a feeder’s effect on the Grade 8 mark is all difficulty, which §3.3 of the main text maintains.

Table 56: **Each way the wedge could be wrong, its check, and where the evidence is**

The wedge could be wrong if	Check	Reported in
<i>Panel A. Standard two-way fixed-effects checks</i>		
schools fall into groups that no student moves between, so their wedges are not on one scale	Coverage of the largest connected set and of the leave-one-out connected set (Kline 2024, 126–27, 151; Kline et al. 2020)	Table 4 of the main text; Appendix S.2 ; Appendix A
with few students on each link between a feeder and a destination, estimation noise inflates the spread of the wedges	Leave-out bias correction of the variance, and a sub-sampling plot (Kline et al. 2020; Kline 2024, 147–52; Bonhomme et al. 2026, 17)	Table 4 of the main text; Appendix S.3 ; Appendix L.1
students bound for different destinations were already on different paths before the move	Marks of students who change school over the four Grade 7 terms, by destination, within the same feeder (Card et al. 2013, 983–86, 991, 996–97; Kline 2024, 141)	Table 4 of the main text; Appendix S.4
the two schools’ effects do not add: particular pairs of feeder and destination, or the change at the move, depart from their sum	Pair effects; mean residuals by wedge decile $\times$ destination decile; the change at the move net of the model’s prediction, by quarter of feeders (Kline 2024, 130, 136–40; Card et al. 2013, 997)	Table 4 of the main text; Appendix S.5
where a student goes depends on something the model does not condition on	Within-feeder permutation test of destinations on observed traits; sorting on the Term 4 mark’s departure from the student’s earlier marks (Card et al. 2013, 989, 992)	Table 4 of the main text; Appendix F ; Appendix S.4
a feeder does not grade all its students by one shift	Mean residuals by the student’s Grade 7 decile $\times$ wedge decile; a slope and shift per Grade 7 quintile (Card et al. 2013, 995–96; Bonhomme et al. 2026, 19)	Table 4 of the main text; Appendix S.5 ; §3.3 of the main text
a school grades some groups of its own students differently	Separate wedges by student group within the same schools (Card et al. 2016, 649–50, 658–61; Bonhomme et al. 2026, 20)	Table 4 of the main text; Appendix S.6
conditioning on the Grade 7 mark stands in poorly for each student’s own effect	A student fixed effect from the eight term marks (Abowd et al. 1999)	Table 4 of the main text; Appendix S.7
a feeder’s wedge drifts across cohorts, so one number per school misleads	Stationarity (Kline 2024, 141; Lachowska et al. 2023)	Table 4 of the main text; Appendix G
<i>Panel B. Specific to school-set marks</i>		
a feeder’s effect on the Grade 8 mark is not all difficulty: the feeder taught skills its marks missed, or its students grow differently after the move	The exclusion restriction (Kline 2024, 140; Bonhomme et al. 2026, 20): maintained; drift within Grade 8 small, not nil	§3.3 of the main text; Appendix D ; Appendix H.8

The wedge could be wrong if	Check	Reported in
the size of a feeder’s wedge carries its intake, through error in the Grade 7 mark or otherwise	The observable placebo; persistence and reliability of the Grade 7 mark; the Monte Carlo	Appendix H.2 ; Appendix N ; §5.1 of the main text
marks are pushed over the promotion threshold	Heaping at the threshold; de-heaping both marks	§2.2 of the main text; §5.2 of the main text; §5.3 of the main text
class teachers mark the same assessment differently	Maintained: the data carry no class identifier	§3.3 of the main text

Notes: Each row states a way the wedge could be wrong, the check that addresses it, and where this paper reports the evidence. Panel A follows the standard checks of the two-way fixed-effects literature (Kline 2024, 140–45; Bonhomme et al. 2026, 16–17); Panel B holds the threats specific to marks that schools set.

What the rows of Table 4 of the main text measure. The noise share is the share of the wedge’s cross-school variance that is estimation noise, by the leave-out correction on the leave-one-out connected set (the schools that stay connected when any one student is dropped), with errors independent across students. With errors shared within a link the share is 8.1% and 7.2% (simulated standard errors 0.8 and 0.8 percentage points, against a precision rule of half a point, which the headline correction meets); weighted by students, as the labour literature weights by person-years, it is 4.2% and 3.9%. Its complement is a signal share of the wedge’s variance; it is not the reliability of the Grade 7 mark. The pre-move row is the change from Grade 7 Term 1 to Term 4, the top quarter of destinations against the bottom quarter, among students from the same primary school, cohorts pooled; the top quarter holds the destinations with the highest destination effects, estimated from the other half of the students or from the other cohort, or, under the mean-mark grouping, the highest raw mean Grade 8 marks. The residual rows use cells of at least 100 students, by wedge decile $\times$ destination decile for the additivity row and by the student’s Grade 7 decile $\times$ wedge decile for the row on one shift per feeder, with deciles weighted by students; the second is net of the Grade 7 margin shared by all feeders. The joint test is a Wald test that the decile interactions are jointly zero. The quintile refit gives each Grade 7 quintile its own slope and shift, the same for every feeder, so each feeder stays one shift; the spread of the wedges changes (SD ratios 0.93 and 0.98) while their order holds. The group rows use schools with at least 20 students in each group (468 and 658 for over-age, 1,031 and 1,202 for gender). The benchmark from noise alone is the slope that noise in the reference group’s wedge would give if both groups shared one wedge: that wedge’s split-half reliability. The population-group contrast is fitted, but on too few schools to read. The student effect is estimated from the eight term marks with persistence held at one, so its school effects are net of each student’s own effect rather than of the Grade 7 mark. Each cohort’s wedges come from one move within months, so drift across cohorts does not enter their identification; the correlation across cohorts is precision-weighted, not corrected for estimation noise.

Samples by row. The pre-move row uses students who change school and have all eight term marks, cohorts pooled (165,408 students on split halves, 144,029 on the other cohort’s estimates, 165,458 on mean marks); the coverage row every province-cohort cell of the National (6 provinces) sample; the sorting row 22,142 feeder-by-cohort cells of the National (6 provinces) sample, cohorts pooled (22,039 for gender and 21,927 for over-age); the stationarity row the 1,280 Gauteng schools in both cohorts; the Panel B row students who change school with Grade 8 Term 1, 2 and 4 marks at one secondary school, cohorts pooled (189,684 students; 176,111 with the other cohort’s wedge), and, for the anchor result, a separate national extract of the same school-based-assessment records (Grade 7 cohorts of 2021 and 2023, seven province-cohort cells; Gauteng not included).

S.1: The model in two-way fixed-effects terms

The model is the two-way fixed-effects design of Abowd et al. (1999) with one compulsory move. Students stand in for workers, and the firm is the school that sets each mark: the feeder in Grade 7, the destination grade cell in Grade 8 (Table 57).

Table 57: **The model here is the two-way fixed-effects model, with one compulsory move and one departure**

Two-way fixed-effects model (labour)	This paper
Worker i	Student i
Firm	The school that sets the mark: feeder j in Grade 7, destination grade cell k in Grade 8
Log wage	The mark, standardised within province and cohort
Worker effect	Latent learning at the end of Grade 7, a_{ij}^7
Firm effect	The school's mark-scale offset, s_j for the feeder and d_k for the destination, estimated as $\alpha_j = \lambda s_j$ and $\gamma_k = -d_k$ (difficulty and possibly teaching)
Job move	The compulsory move from Grade 7 to Grade 8, once, at one point, by rule, for promoted students; students at combined schools change grader without changing school
The one departure	Learning need not carry over one for one: the slope β_1 on the Grade 7 mark is left free, and it estimates $\lambda\rho$, persistence times the Grade 7 mark's reliability (Appendix N)

Notes: The left column names the objects of the two-way fixed-effects model of Abowd et al. (1999) as labour economics uses it; the right column gives their counterparts in the model of §3 of the main text, in its notation.

Identification rests on the two conditions of the AKM literature, exogenous mobility and additivity, within a connected set (Kline 2024, 126); §3.3 of the main text states them for this paper. With no student effect, exogenous mobility asks more than in AKM, which allows sorting on each person's own fixed effect (Card et al. 2013, 989); the student-fixed-effect version ([Appendix S.7](#)) allows that sorting but imposes persistence of one. Because the rule sets the move, its timing says nothing about the student: no move follows a bad term, and the marks of students bound for different destinations ran nearly parallel before it ([Appendix S.4](#)). And each student moves once, so a mark cannot carry a history of moves, as a wage can carry a worker's history of outside offers (Postel-Vinay and Robin 2002, 2297). Where a student goes is still chosen. A few students made a second move that the rule did not fix, changing school within Grade 7 or within Grade 8 before the Term 2 mark: 0.7 and 1.1 per cent of the 2018 and 2022 Gauteng estimation samples. They stay in the sample; Kramarz et al. (2015, 580) find that such moves bias a school's effect only where they make up most of its entrants. Only students who enter Grade 8 are linked, and the feeder's own marks decide promotion, so the estimation sample is selected on the marks being calibrated (§2 of the main text). At a combined school the feeder and the destination grade cell belong to one institution, so its stayers compare the school's Grade 7 marking with its own Grade 8 marking.

S.2: The connected set

Wedges are comparable only within a set of schools that students' moves link, so each province-cohort cell is estimated on its largest connected set (Kline 2024, 126–27). In every cell of Table 58, that set holds at least 98.8 per cent of the primary schools and 99.3 per cent of the linked students.

Table 58: **Almost every primary school sits in the largest connected set of its cell**

Province	All primary schools	Largest connected set			Components
		Primary schools	Primary schools (%)	Students (%)	
<i>Panel A. 2018 cohort</i>					
Eastern Cape	3,117	3,079	98.8	99.3	37
Gauteng	1,432	1,424	99.4	99.8	9
KwaZulu-Natal	2,543	2,535	99.7	99.8	9
Limpopo	2,151	2,143	99.6	99.9	9
Mpumalanga	970	964	99.4	99.8	7
North West	834	833	99.9	100.0	2
<i>Panel B. 2022 cohort</i>					
Eastern Cape	3,142	3,129	99.6	99.8	13
Gauteng	1,484	1,479	99.7	99.9	6
KwaZulu-Natal	3,069	3,062	99.8	99.9	8
Limpopo	2,240	2,236	99.8	100.0	5
Mpumalanga	1,050	1,047	99.7	100.0	4
North West	939	937	99.8	99.9	3

*Notes: Each row is one province-cohort cell, and the model is estimated separately in each. A primary school is in a connected set when the moves of students link it, directly or through other schools, to every other school in the set; wedges are comparable only within a set, so each cell is estimated on its largest connected set and the schools outside it are dropped. Primary schools are those with students linked to a Grade 8 destination. The shares are of the cell's primary schools and of its linked students; components count the separate sets in the cell's network of moves. No significance star appears in this table; where stars appear elsewhere in this document they follow one ladder: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an em dash. None arises here. Source: the display record outputs/tables/mathematics/tab_akm_appendix_coverage.csv, built from outputs/tables/mathematics/connected_set_coverage.csv; its numbers and selectors are in outputs/tables/mathematics/tab_akm_appendix_sources.csv. Sample: the National (6 provinces) sample, Mathematics, the 2018 and 2022 cohorts: every province-cohort cell, with 533,419 and 665,514 linked students and 11,047 and 11,924 primary schools.*

The leave-out correction of [Appendix S.3](#) needs more: the leave-one-out connected set, the schools that stay connected when any one student is dropped (Kline et al. 2020; Kline 2024, 151). In Gauteng that set keeps almost every connected primary school (the counts in [Table 59](#) against [Table 58](#)).

S.3: Limited mobility

With few students on each link between a feeder and a destination, estimation noise inflates the spread of the wedges: limited-mobility bias. [Table 59](#) reports the standard leave-out correction of the variance (Kline et al. 2020; Kline 2024, 147–52; Bonhomme et al. 2026, 17) in four versions, each with a simulation check.

Table 59: **Estimation noise is a small share of the variance of the wedge, but not nil**

Correction and weighting	Primary schools	SD of the wedge		Noise share (%)	Precision
		Plug-in	Corrected		

Panel A. 2018 cohort

(continued)

Correction and weighting	Primary schools	Plug-in	Corrected	Noise share (%)	Precision
Errors independent across students; each primary school counts once	1,421	0.471	0.458	5.4 (0.4)	Pass
Errors independent across students; schools weighted by students	1,421	0.446	0.436	4.2 (0.2)	Pass
Errors shared within a link; each primary school counts once	1,409	0.471	0.452	8.1 (0.8)	Fail
Errors shared within a link; schools weighted by students	1,409	0.446	0.432	6.0 (0.7)	Fail
<i>Panel B. 2022 cohort</i>					
Errors independent across students; each primary school counts once	1,472	0.468	0.455	5.5 (0.3)	Pass
Errors independent across students; schools weighted by students	1,472	0.448	0.439	3.9 (0.1)	Pass
Errors shared within a link; each primary school counts once	1,461	0.468	0.451	7.2 (0.8)	Fail
Errors shared within a link; schools weighted by students	1,461	0.448	0.436	5.2 (0.5)	Pass

*Notes: Each row is a leave-out correction of the variance of the wedge across primary schools (Kline et al. 2020): the plug-in variance of the estimated wedges, less the part that estimation noise contributes. The noise share is that part as a share of the plug-in variance, and the first row of each panel is the headline. With errors independent across students, the correction leaves out one student at a time, on the leave-one-out connected set (the schools that stay connected when any one student is dropped); with errors shared within a link, it leaves out at once all the students on one link between a feeder and a destination, on the corresponding leave-one-link-out set, so errors may be correlated within a link. The primary-school count is that of each row's leave-out set. Where each primary school counts once, the weighting is that of the wedge's SD in the main text; where schools are weighted by students, each counts by its students, as the labour literature weights by person-years. SDs are of the wedge across primary schools, in within-province standard deviations (SD) of the mark. The plug-in SD is the full fit's, and the corrected SD is the plug-in SD times the square root of one minus the noise share. The simulation draws 2,000 outcomes on each row's own design, with the fit's estimates as the truth and errors independent across students. The correction passes the bias check in every row: the 95% interval of its mean error lies within 1% of the true variance. Precision passes when the standard deviation of the relative error across draws is at most 3% and that of the noise share at most 0.5 percentage points; the latter, the simulated standard error of the noise share, is in parentheses beneath it. No significance star appears in this table; where stars appear elsewhere in this document they follow one ladder: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an*

em dash. None arises here. Source: the display record outputs/tables/mathematics/tab_akm_appendix_noise.csv, built from analyses/wedge_noise_correction/outputs/tables/corrected_variance.csv and validation.csv; its numbers and selectors are in outputs/tables/mathematics/tab_akm_appendix_sources.csv. Sample: Gauteng Mathematics, the 2018 and 2022 cohorts estimated separately on the published estimation sample (106,514 and 135,852 students). Each correction runs on its own leave-out set: 106,414 and 135,668 students leaving out one student, 105,702 and 135,087 leaving out one link.

The share of the wedge’s variance that is estimation noise is small, but not nil. In the headline row (independent errors, each school counted once), the corrected standard deviation sits a little below the plug-in value the main text reports (0.471 within-province standard deviations in 2018). Weighting by students shrinks the share. Allowing errors shared within a link enlarges it, but this robustness correction, unlike the headline, fails the precision rule in 2018 and, with schools counted once, in 2022. The share’s complement is a signal share of the wedge’s variance, not the reliability of the Grade 7 mark.

A sub-sampling plot (Bonhomme et al. 2026, 16) asks whether that answer depends on how many leavers each primary has. Each draw keeps the same random share of every primary’s leavers, from a tenth to three-quarters, refits the model on the leave-one-out connected set and compares the variance of the wedge with the full data’s corrected variance over the same primaries (Figure 29). As primaries thin, the plug-in variance climbs: with a tenth of the leavers, about 8 per primary in 2018 and 9 in 2022, it is 1.61 and 1.48 times the reference. The corrected variance stays within two standard errors of the reference at every share in both cohorts, and never exceeds 1.05 times it. The correction removes the noise that thin primaries add, so the corrected spread of the wedges does not depend on how many leavers a primary sends. Every corrected point sits at or slightly above the reference, as shocks shared by the students on one link would make it; the version of the plot that allows such shocks is not run.

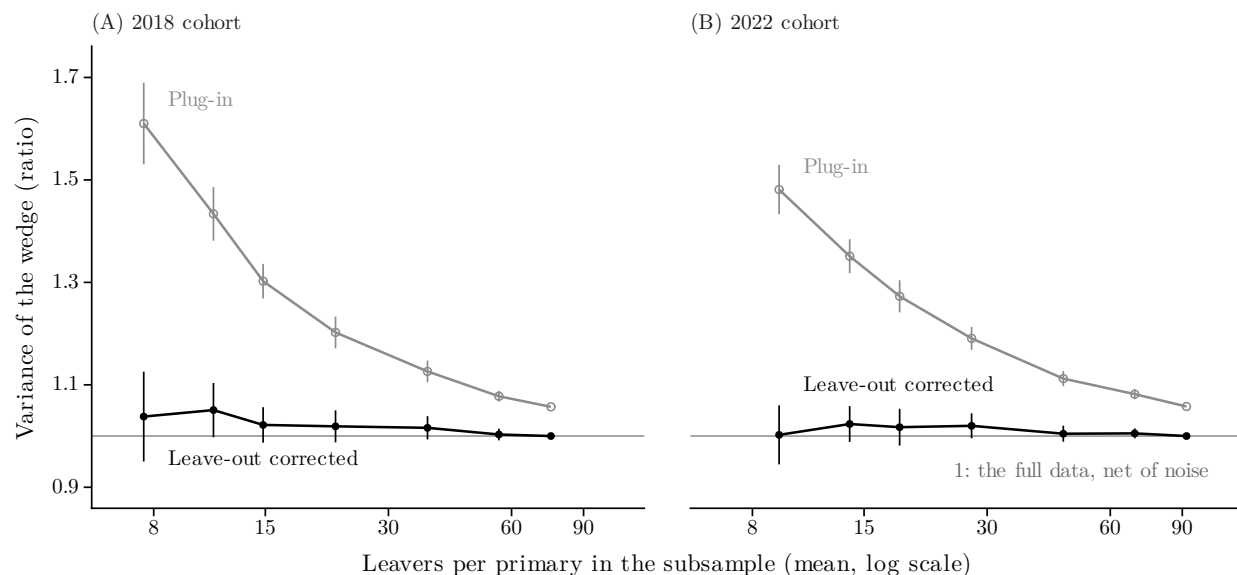


Figure 29: Thin primaries inflate the plug-in variance, not the corrected one

Notes: Each draw keeps the same random share (0.1, 0.15, 0.2, 0.3, 0.5, 0.75) of every primary’s leavers, refits the model on the leave-one-out connected set and divides the variance of the wedge across primaries (each primary counted once) by the full data’s leave-out corrected variance over the same primaries. Open circles: plug-in; filled: leave-out corrected (Kline et al. 2020), errors independent across students. Points are means over ten draws; whiskers are two standard errors across draws. The horizontal axis is on a log scale; the right-most points are the full data. The missing-value glyph for this document is an em dash. None arises here. Source: analyses/wedge_noise_correction/outputs/tables/subsample_summary.csv. Sample: Gauteng, Mathematics, 2018 and 2022 cohorts, the published fit.

A minimum number of students per link, a common remedy for limited mobility, adds noise here rather than removing it. Dropping every link between a feeder and a destination that carries fewer than five students removes 19.9% of the 2018 sample and 18.2% of the 2022 sample by the rule, and on the feeders both fits share it widens the standard deviation of the wedge by a factor of 1.20 (2018) and 1.19 (2022). Split-half estimates locate the rise. Estimation noise grows from 3.6% of the wedge’s variance to 16.6% in 2018, and from 3.9% to 15.8% in 2022. Net of that noise the two fits order feeders almost alike, with correlations of 0.95 and 0.99, although the thin-link wedges are somewhat more spread (by a factor of 1.12 and 1.12 in standard deviation). Fewer students alone do not add the noise: removing as many students from each feeder at random, while keeping every link, leaves the standard deviation where it was (a factor of 0.99 and 0.99). The noise comes from the links the rule cuts, which tie the destination effects together, and the baseline keeps them all.

Table 60 gives the decomposition the two-way fixed-effects literature reports first (Abowd et al. 1999; Kline 2024, 146): the variance of the Grade 8 mark split into the persistence term, $\hat{\beta}_1 g_{ij}^7$, the wedge, the destination effect, twice each covariance and the residual, each weighted by students and corrected for estimation noise by the same leave-out method. The wedge accounts for 19.1 per cent of that variance in 2018 and 19.2 per cent in 2022, against 36.1 and 35.0 per cent for the persistence term. The persistence term and the wedge covary negatively: students with high Grade 7 marks tend to come from primaries with low wedges, that is, primaries with easier assessments. The wedge and the destination effect barely covary (student-weighted correlations of 0.09 and -0.06), so students from primaries with harder assessments do not gather at destinations with high effects.

Table 60: **The wedge explains much of the Grade 8 mark and barely covaries with destinations**

Share of the variance (%)	2018		2022	
	Plug-in	Corrected	Plug-in	Corrected
Persistence term	36.1	36.1	35.0	35.0
Wedge	19.9	19.1	20.0	19.2
Destination effect	24.0	23.3	20.4	19.8
$2 \times \text{Cov}(\text{persistence, wedge})$	-16.2	-16.2	-9.8	-9.8
$2 \times \text{Cov}(\text{persistence, destination})$	1.5	1.5	-1.0	-1.0
$2 \times \text{Cov}(\text{wedge, destination})$	3.2	4.0	-3.0	-2.3
Residual	31.5	32.2	38.5	39.1
Variance of the Grade 8 mark (SD^2)	0.999	0.999	0.998	0.998
Observations	106,414	106,414	135,668	135,668

*Notes: Shares of the variance of the Grade 8 Term 2 mark, in per cent, from the two-way model of the main text; each term is weighted by students. Plug-in columns use the estimated coefficients; corrected columns subtract the leave-out estimate of estimation noise in each quadratic form (Kline et al. 2020), errors independent across students; the corrected residual is the remainder. A simulation on the estimation sample’s own design, with the fit as truth, finds no bias in any corrected term. Marks are in within-province standard deviations (SD), so the variance row is in SD squared. No significance star appears in this table; where stars appear elsewhere in this document they follow one ladder: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an em dash. None arises here. Source: the display record outputs/tables/mathematics/tab_akm_appendix_decomposition.csv, built from analyses/wedge_noise_correction/outputs/tables/variance_decomposition.csv and variance_decomposition_validation.csv; its numbers and selectors are in outputs/tables/mathematics/tab_akm_appendix_sources.csv. Sample: Gauteng Mathematics, the 2018 and 2022 cohorts estimated separately on the published estimation sample: the leave-one-out connected set of the published fit (106,414 and 135,668 students).*

S.4: Marks before the move

Figure 30 plots the marks of students who change school, term by term, grouped by quarters of feeders and of destinations (Card et al. 2013, 983–86, 991; Kline 2024, 141). If students do not split across destinations

on their growth in learning, those from one feeder bound for different destinations share a path before the move. The transition rule fixes that move for promoted students: once, between Grade 7 Term 4 and Grade 8 Term 1, though not where it leads (Kramarz et al. 2015, 577).

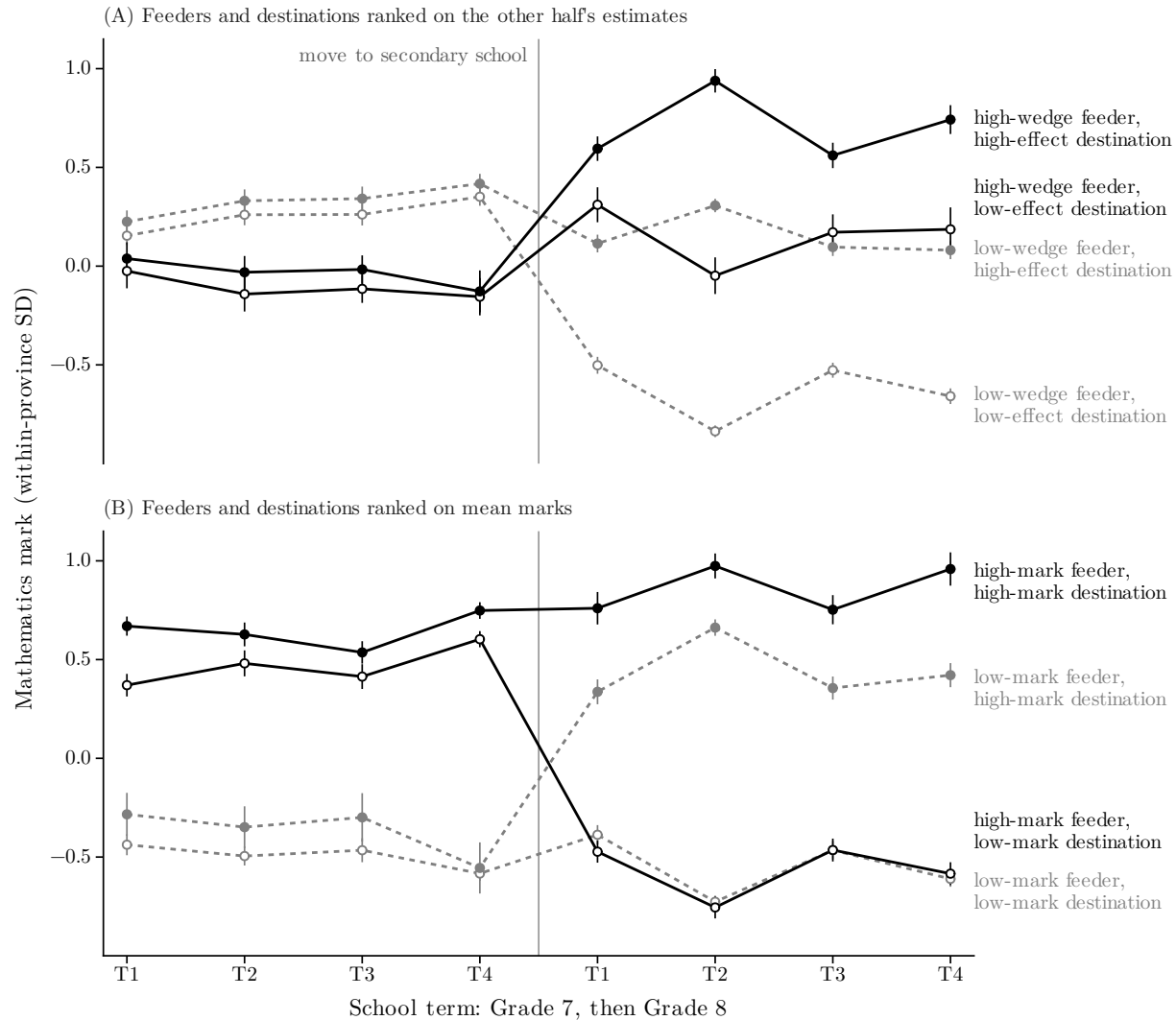


Figure 30: Paths to high- and low-effect destinations ran nearly parallel before the move, by feeder quarter

Notes: Each line follows the mean Mathematics mark, term by term, of one group of students who change school between Grade 7 and Grade 8, with 95% intervals from standard errors clustered by primary school. Students are grouped by the quarter of feeders they leave and the quarter of destinations they enter, both quarters weighted by students within each cohort, and the figure draws the four corner groups: solid black lines for the top quarter of feeders and dashed grey lines for the bottom quarter, filled markers for the top quarter of destinations and open markers for the bottom quarter. In panel A, feeders are ranked on the wedge and destinations on the destination effect, each estimated on the other random half of the students, so a student's own marks never set the student's group; the top quarter of feeders has the largest wedges, and the top quarter of destinations the largest destination effects. In panel B, feeders are ranked on their mean raw Grade 7 mark and destinations on their mean raw Grade 8 mark, each leaving the student's own mark out, the grouping the labour literature uses. Marks are standardised within cohort and term, in within-province standard deviations (SD), so a shift common to every group from one term to the next reflects that term's assessment and only the differences between groups carry information. Each

line pools all the feeders of one quarter, so its path before the move also reflects differences between those feeders; the pre-move contrast in the main text compares destination quarters among students of the same primary school. The vertical line marks the move, between Grade 7 Term 4 and Grade 8 Term 1. The corner groups' changes at the move differ because each quarter of destinations holds a different mix of schools for students of different feeders; net of what the same model, fitted to the Grade 8 Term 1 mark, predicts for the same students, the gain from entering a top rather than the bottom quarter of destinations is the same for every quarter of feeders, as the additivity checks in this appendix report. Source: *analyses/seminar_checks/outputs/tables/tabQ05_groups.csv* (panel A) and *tabQ05_groups_labour.csv* (panel B), the rows for the pooled cohorts. Sample: Gauteng Mathematics, the 2018 and 2022 cohorts pooled: students in the published estimation sample who change school between Grade 7 and Grade 8 and have all eight term marks, at the same school in every term of each grade. Panel A's groups are drawn from 165,408 students and panel B's from 165,458; the four corner groups drawn hold 40,601 and 42,933 students.

With destinations ranked on the destination effect, the change from Grade 7 Term 1 to Term 4 within a feeder barely differs between students bound for the quarter of destinations with the highest effects and the quarter with the lowest (Card et al. 2013, 996–97), on split-half groups and on the other cohort's estimates (Table 4 of the main text). With schools ranked on mean marks, as in the standard labour diagnostic (Kline 2024, 141), students bound for high-mark destinations gain a few hundredths of an SD more before the move than their schoolmates bound for low-mark destinations. The Grade 8 terms are shown for completeness. The lines part by unequal amounts at the move because each quarter of destinations holds a different mix of schools for students of feeders with harder and easier assessments; net of that mix, the change at the move does not depend on the feeder (Appendix S.5). The groups above are not held at one Term 4 mark, the mark the model conditions on. A sharper version of the question (Card et al. 2013, 992) compares students of one feeder with the same Term 4 mark and asks whether those whose Term 4 mark ran above their own Term 1 to Term 3 average enter different destinations. In 2018 they do, slightly: each SD of that departure lowers the destination's effect by 0.013 SD (standard error 0.004, clustered by feeder), against a rise of 0.012 SD per SD of the Term 4 mark itself. Added to the model as a control, after removing its correlation with the Term 4 mark within each feeder, the departure can move a wedge only through this sorting. The wedges then correlate with those of the model without it at 0.998 (Spearman), and their variance is 4.6 per cent larger. In 2022 the Term 4 mark is built from the three earlier terms, so the departure barely predicts the Grade 8 mark and the wedges do not move; the test has power only in 2018.

S.5: Additivity

Additivity can fail in two ways (Kline 2024, 130; Card et al. 2013, 995–97; Bonhomme et al. 2026, 19): a pair of feeder and destination whose joint effect departs from the sum of their effects, or a feeder that does not grade all its students by one shift. Table 61 and Figure 31 report both checks.

Table 61: **The two school effects nearly add; quintile slopes leave the order of feeders almost intact**

Statistic	2018	2022
<i>Panel A. Pairs of feeder and destination</i>		
Adjusted R^2 , one effect per feeder and per grade cell	0.678	0.609
Adjusted R^2 , a separate effect per pair	0.697	0.629
SD of pair effects about the sum of the two schools' effects, net of noise, links of two or more students (SD)	0.079	0.078
Largest absolute cell mean, wedge decile $\times$ destination decile (SD)	0.039	0.037
Joint test that the cell means are zero, p-value	0.020	0.194
<i>Panel B. One shift per school</i>		
Largest absolute cell mean, Grade 7 decile $\times$ wedge decile, net of the Grade 7 margin (SD)	0.369	0.374
Largest absolute cell mean, Grade 7 decile $\times$ destination decile, net of the Grade 7 margin (SD)	0.248	0.105

(continued)

Statistic	2018	2022
<i>Panel C. A slope and shift per Grade 7 quintile, shared by all feeders</i>		
Rank correlation with the baseline wedge (Spearman)	0.997	0.999
SD of the refit wedges over SD of the baseline wedges	0.93	0.98
Students	106,514	135,852
Primary schools	1,424	1,479

*Notes: Panel A compares the model of the main text, one effect for each primary school and one for each destination grade cell, with a model that gives each pair of feeder and destination its own effect, the standard deviation of those pair effects about the sum of the two schools' effects, corrected for estimation noise by the leave-out estimator of Kline, Saggio and Sølvesten (2020) on the links that carry two or more students, and reports the largest absolute mean residual over the cells of wedge decile $\times$ destination decile, with a Wald test that the decile interactions are jointly zero, clustered by primary school. Panel B reports the largest absolute mean residual by the student's Grade 7 decile $\times$ wedge decile, and by Grade 7 decile $\times$ destination decile, each net of the Grade 7 margin shared by all schools: each cell's mean less the mean of its Grade 7 decile over all schools. Panel C refits the model with its own slope and shift for each Grade 7 quintile, the same for every feeder, so each feeder keeps one shift: it tests the form of the Grade 7 slope, not a slope per feeder. Residuals are in within-province standard deviations (SD) of the mark. Cells hold at least 100 students, and deciles of schools are weighted by students; wedge decile 1 holds the feeders with the easiest assessments, destination decile 1 the lowest destination effects and Grade 7 decile 1 the lowest marks. Not run here: the change in marks at the move, from Grade 7 Term 4 to Grade 8 Term 1, net of which students make each move; and the order of feeders under a slope per feeder. No significance star appears in this table; where stars appear elsewhere in this document they follow one ladder: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an em dash. None arises here. Source: the display record outputs/tables/mathematics/tab_akm_appendix_additivity.csv, built from analyses/seminar_checks/outputs/tables/tabQ04_match.csv, tabQ04_summary.csv, tabQ07_within_pair.csv and tabQ10_links.csv, and analyses/wedge_noise_correction/outputs/tables/match_summary.csv; its numbers and selectors are in outputs/tables/mathematics/tab_akm_appendix_sources.csv. Sample: Gauteng Mathematics, the 2018 and 2022 cohorts estimated separately on the published estimation sample (106,514 and 135,852 students; 1,424 and 1,479 primary schools), combined-school students included.*

A separate effect for each pair adds little fit, and mean residuals by wedge decile and destination decile are small (panels A and B of Figure 31), but these feeder–destination interactions are jointly significant in 2018, though not in 2022 (Table 61). Net of estimation noise, the pair effects depart little from the sum of the two schools' effects. On the links that carry two or more students (88 and 90 per cent of students), the leave-out correction of Kline et al. (2020) puts the SD of that departure at 0.079 in 2018 and 0.078 in 2022, against a wedge SD of 0.471 and 0.468, so the two schools' effects account for 98.7 and 98.3 per cent of the noise-free variance of the pair effects (Kline 2024, 136–40). The departure includes any shock that students on one link share, such as a Grade 8 class, so difficulty specific to a pair of schools is likely smaller still. A student alone on a link, as 10,952 are in 2018 and 12,558 in 2022, is fitted exactly by the pair model, and the noise in that student's mark cannot be measured. Over all links the SD therefore rests on an assumption about it: 0.134 in 2018 if those students' marks are as noisy as other students', 0.079 if all of their departure is noise and 0.218 if none of it is (0.139, 0.079 and 0.229 in 2022). At the move itself, from Grade 7 Term 4 to Grade 8 Term 1, the two schools' effects also add once the schools in each group are allowed for. Among students who change school, grouped on estimates from the other half of the students, entering the quarter of destinations with the highest rather than the lowest destination effects raises the change at the move by 0.55 SD for students of the quarter of feeders with the easiest assessments but by only 0.26 SD for those with the hardest, a gap that a test of equal contrasts finds in each of ten random splits. Who moves where accounts for the gap: each quarter of destinations holds a different mix of schools for students of different feeders, and the same model fitted to the Term 1 mark predicts the gap from those schools. Net of that prediction for the same students, every contrast lies within 0.02 SD of zero and none differs by quarter of feeders (smallest p-value over the ten splits 0.19, and 0.14 when the prediction comes from the other half of the students). On the other cohort's estimates the raw contrasts already agree, and the net ones do too

(p-value 0.36).

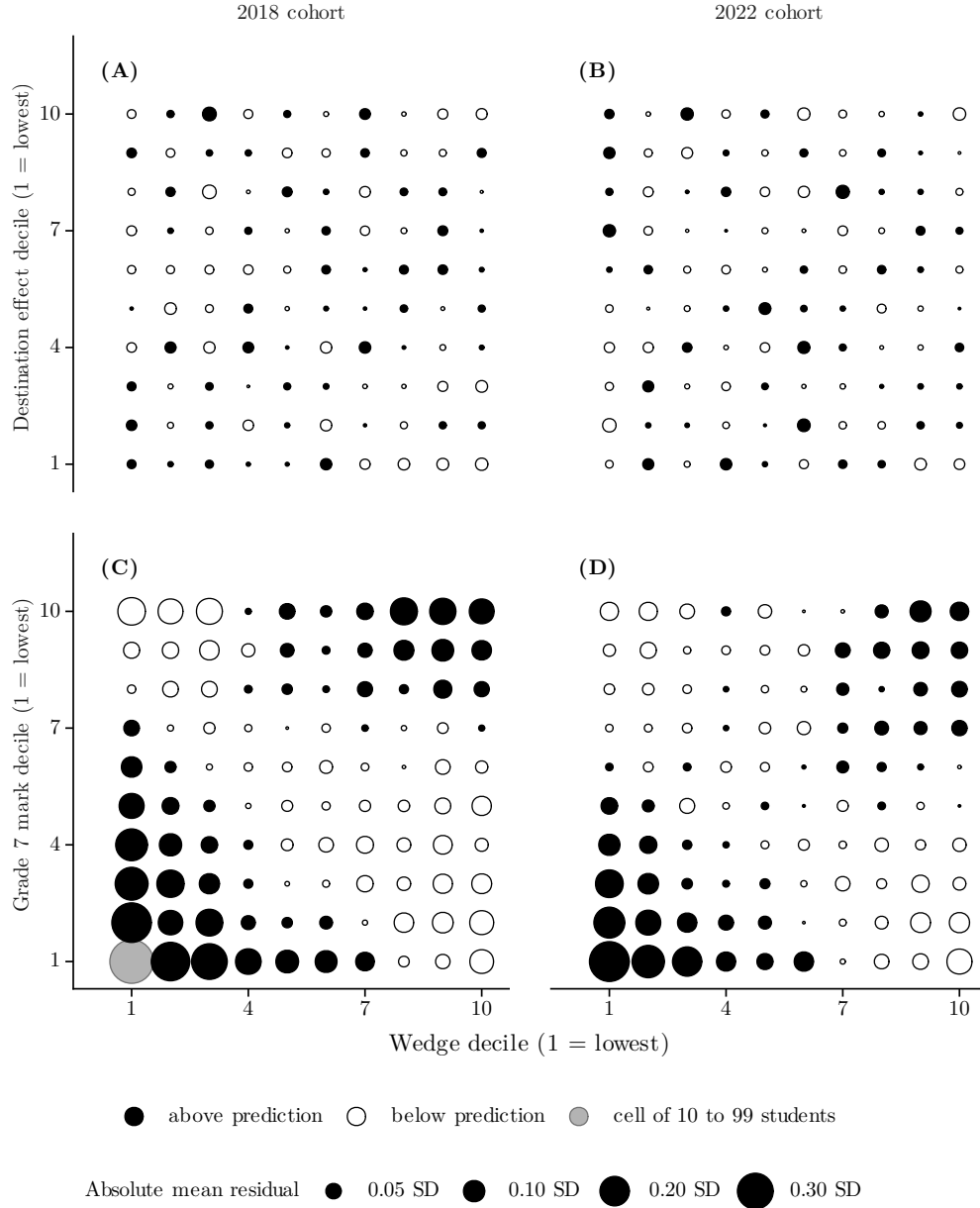


Figure 31: Residuals are small across pairs of schools but vary with student marks within a feeder

Notes: Each circle is the mean residual of the model of the main text in one cell, in within-province standard deviations (SD) of the mark: its area is proportional to the absolute mean, filled where the cell's students score above the model's prediction and open where they score below. Under the model, with each school one shift, every cell's mean residual is zero in expectation. Panels A and B group students by the decile of their feeder's wedge and the decile of their destination's effect. Panels C and D group them by the wedge decile and the decile of their own Grade 7 mark, net of the Grade 7 margin shared by all feeders: each cell's mean less the mean of its Grade 7 decile over all feeders, weighted by students. Deciles of schools are weighted by students; wedge decile 1 holds the feeders with the easiest assessments, destination decile 1 the lowest destination effects and Grade 7 decile 1 the lowest marks.

Grey circles are cells of 10 to 99 students (1 in all); cells under 10 students would not be drawn, and none arises. Source: analyses/seminar_checks/outputs/tables/tabQ04_cells_ag.csv (panels A and B) and tabQ04_cells_zs.csv (panels C and D). Sample: Gauteng Mathematics, the 2018 and 2022 cohorts estimated separately on the published estimation sample (106,514 and 135,852 students), combined-school students included; cells of at least 100 students: 100 and 100 in panels A and B, 99 and 100 in panels C and D.

Within a feeder, the shift varies with the student's Grade 7 mark (panels C and D): low-mark students of feeders with easier assessments and high-mark students of feeders with harder ones score above the model's prediction, the opposite pairs below it. Net of the Grade 7 margin, low-mark students at the feeders with the easiest assessments score 0.369 SD above prediction in 2018 (0.374 in 2022), against a cross-school SD of the wedge of 0.471. A slope and shift per Grade 7 quintile, shared by all feeders, barely reorders them (rank correlations of 0.997 and 0.999 with the baseline; Table 61). A slope for each feeder takes up most of the pattern: the largest net cell falls to 0.154 SD in 2018 and 0.175 in 2022. The feeder slopes average 0.60 and, net of sampling error, differ across feeders with a standard deviation of 0.19 in 2018 and 0.14 in 2022, rising from 0.50 in the fifth of feeders with the easiest assessments to 0.71 in the fifth with the hardest in 2018. Each school's calibrated level stays where it was, and so, nearly, does the order of wedges read at the average mark, but a feeder's wedge for students at the 10th percentile of the Grade 7 mark and its wedge for students at the 90th correlate only 0.61 in 2018 and 0.71 in 2022, net of sampling error. A proportional grading scale would steepen the slope of a feeder with harder assessments by narrowing its Grade 7 marks relative to its students' Grade 8 marks. The ratio of the two spreads contributes little to the steeper slope in 2018 and works against it in 2022; the steeper slope comes instead from a closer correlation between the two marks, and most of it disappears once the term-to-term reliability of the feeder's Grade 7 marks and the spread of its students' Grade 8 marks are held fixed. At the destination, the Grade 8 mark rises more steeply with the Grade 7 mark at secondaries with high destination effects than at those with low ones: at the decile of destinations with the lowest destination effects, top-decile students score 0.248 SD below prediction in 2018, and bottom-decile students 0.105 SD above it in 2022, net of the Grade 7 margin. Each secondary's slope differs from the common one, with a pupil-weighted SD of 0.18 and 0.14 around a mean near 0.58, and correlates 0.43 across the two cohorts. Letting the slope differ by secondary, and scale the feeder's offset with it as case (i) of [Appendix O.10](#) implies, leaves rank correlations of 0.981 and 0.994 with the baseline wedges and their spread 6 and 3 per cent narrower.

Part of the mark dependence is common to every feeder. Among students who leave the same primary for the same secondary school, the Grade 8 mark bends away from a single line fitted within each pair of schools at both ends of the Grade 7 range: students below 30% score 0.13 SD above it in 2018 and those at 80% or more 0.14 SD (0.21 and 0.09 in 2022), and across the six provinces and two cohorts the two ends lie 0.11 to 0.21 and 0.08 to 0.20 SD above it. Threshold heaping does not produce the bend: dropping every Grade 7 mark within two of 30, 40 or 50, and every zero, leaves it at least as wide. Nor does a floor in the Grade 8 mark: 16 per cent of students below 30% in Grade 7 score below 10% in Grade 8. A single line through two term marks at the same school bends as much or more, between Terms 3 and 4 of Grade 7 at the primary and between Terms 2 and 4 of Grade 8 at the secondary, so the bend belongs to school marks at the extremes of their scale rather than to learning over the transition. A separate intercept and slope for each of the seven achievement levels, shared by all feeders, raises the slope from 0.27 below 30% to 1.00 at 80% or more in 2018, leaves the order of feeders unchanged (Spearman at least 0.996) and narrows the wedges' spread to 0.92 of its baseline in Gauteng in 2018.

S.6: Groups of students within schools

If a school grades two groups of its students on one scale, wedges estimated on each agree up to noise (Card et al. 2016, 649–50, 658–61; Bonhomme et al. 2026, 20). Figure 32 plots wedges estimated on over-age against on-age students, and on girls against boys; the contrast by population group is fitted, but too few primary schools teach enough students of both groups to read it.

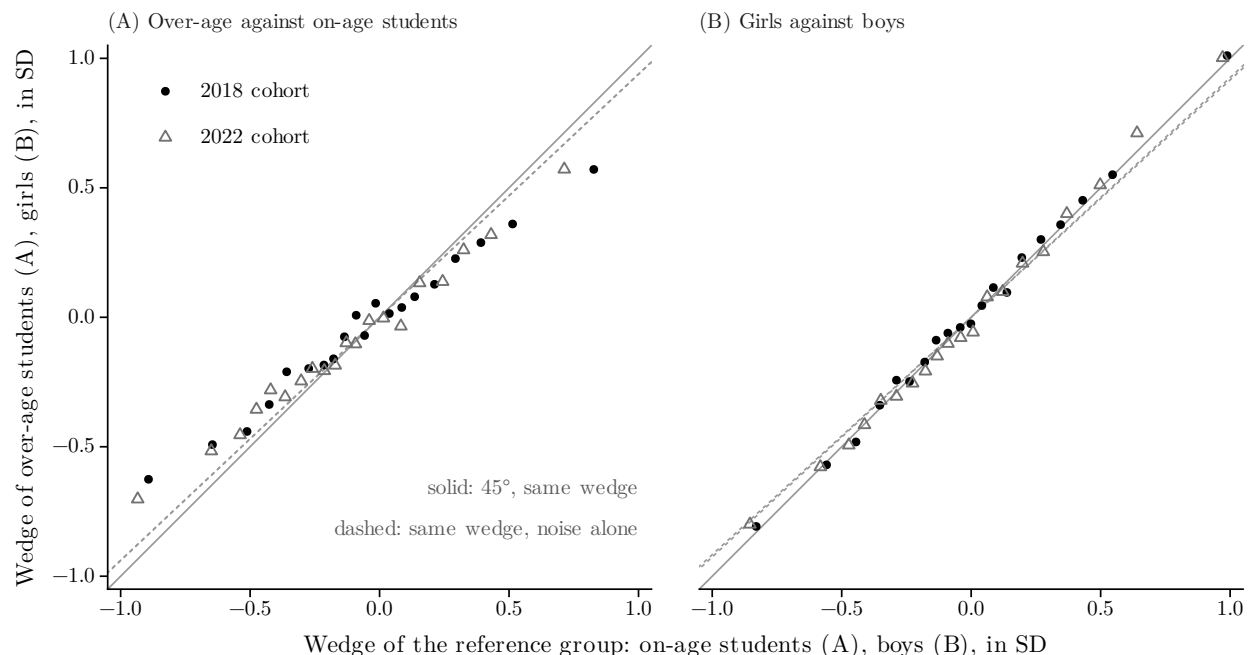


Figure 32: **Over-age students' wedges are narrower by their flatter Grade 7 slope; girls' slightly wider**

Notes: Each point is the mean wedge of the primary schools in one of 20 bins of the reference group's wedge, each bin holding about the same number of schools: on-age students are the reference group in panel A and boys in panel B. Each group's wedge comes from its own two-way fixed-effects fit on its own connected set and is centred within that fit; schools enter with at least 20 students of each group. Wedges are in within-province standard deviations (SD) of the mark. The solid line is the 45° line, on which the two groups' wedges agree. The dashed lines, one for each cohort, have slope equal to the split-half reliability of the reference group's wedge, the slope that estimation noise alone would give if both groups shared one wedge. The population-group contrast is fitted, but on too few schools to read. Source: analyses/seminar_checks/outputs/tables/tabQ09_bins.csv; the reliabilities from tabQ09_groups.csv. Sample: Gauteng Mathematics, the 2018 and 2022 cohorts estimated separately on the published estimation sample: the primary schools with at least 20 students of each group, 468 and 658 for over-age, 1,031 and 1,202 for gender.

Wedges estimated on over-age students are narrower than those on on-age students, lying on a flatter line than noise alone would give (panel A). Net of sampling error, the over-age wedge rises 0.76 SD per SD of the on-age wedge in 2018 and 0.82 in 2022, close to the ratio of the over-age to the on-age Grade 7 slope (0.75 and 0.80). This is what one grading scale predicts. A school that raises every student's Grade 7 mark by the same amount lowers each group's wedge by that amount times the group's own slope, so the group whose marks predict Grade 8 less closely has the narrower wedge. Divided by each group's slope, the two wedges nearly agree: the over-age wedge then rises 1.02 and 1.04 per unit of the on-age wedge. Over-age students' slope is flatter because, within a school, the part of their Grade 7 marks that persists from term to term is less spread out and the part that does not is larger. The ratio of their Grade 8 spread to their Grade 7 spread matches on-age students', so their marks are not set on a narrower scale. Their lower marks explain part of the flatter slope: a slope per feeder, read at each group's mean mark, would give the over-age wedge a slope of 0.87 and 0.95 on the on-age wedge (Appendix S.5). Wedges estimated on girls are slightly wider than those on boys (panel B): net of sampling error, the girls' wedge rises 1.09 SD per SD of the boys' wedge in 2018 and 1.11 in 2022. The difference in width needs no separate scale for girls; it comes through two channels. Within a school, girls' Grade 7 marks track their Grade 8 marks more closely than boys' do (correlations of 0.62 and 0.57 in 2018), so girls' slope is steeper, and divided by each group's own slope the girls' wedge rises 1.05 and 1.05 per unit of the boys'. Girls also enter with Grade 7 marks 0.29

within-province SD above boys' at the same school in 2018 and 0.38 in 2022, where a feeder's wedges fan out ([Appendix S.5](#)). A model with a slope per feeder, a shift for girls at each feeder and one destination effect for both sexes reads the two wedges at the same mark. There the girls' wedge rises 0.97 and 0.98 SD per SD of the boys', and the girls' shift varies across schools with a standard deviation 0.15 and 0.14 times the wedge's, net of sampling error. Giving each destination its own effect for girls as well moves the slope to 1.02 and 1.06. Table 4 of the main text prints the slopes and their noise benchmarks.

S.7: A student fixed effect

The student-fixed-effect version fits each student's eight term marks with the student's own effect, the effect of the school that set each mark and a grade-by-term effect (Abowd et al. 1999): the opposite trade to the baseline, allowing sorting on that effect but imposing persistence of one.

Table 62: **A student fixed effect orders feeders much as the wedge in 2022, less in 2018**

Statistic	2018	2022
Rank correlation with the baseline wedge (Spearman)	0.876	0.949
Correlation with the baseline wedge (Pearson)	0.882	0.949
SD of the student-fixed-effect version's feeder effects	0.563	0.581
SD of the baseline wedges	0.471	0.468
Primary schools	1,424	1,479
Students	106,514	135,852
Term marks	817,988	1,026,096

*Notes: The student-fixed-effect version models each of a student's eight term marks (Grade 7 and Grade 8, Terms 1 to 4) as the student's own effect, plus the effect of the school that set the mark (the feeder in Grade 7, the destination grade cell in Grade 8), plus a grade-by-term effect (Abowd et al. 1999). Learning therefore carries over one for one (persistence held at one), and the school effects are net of each student's own effect rather than of the Grade 7 mark; the two sets of feeder effects are different objects by construction. The student-fixed-effect version's feeder effect is signed so that a positive value means harder assessments, as the wedge is, and both are centred over the fit's primary schools. Correlations are across the primary schools in both fits; rank correlations are Spearman's and the other is Pearson's. SDs are across primary schools. The student-fixed-effect version's feeder effects are in within-province SD of Grade 7 term marks and the baseline wedges in SD of Grade 8 marks, so their spreads cannot be compared directly. No significance star appears in this table; where stars appear elsewhere in this document they follow one ladder: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The missing-value glyph for this document is an em dash. None arises here. Source: the display record outputs/tables/mathematics/tab_akm_appendix_student_fe.csv, built from analyses/seminar_checks/outputs/tables/tabQ14_compare.csv and tabQ14_sample.csv; its numbers and selectors are in outputs/tables/mathematics/tab_akm_appendix_sources.csv. Sample: Gauteng Mathematics, the 2018 and 2022 cohorts estimated separately on the published estimation sample: 106,514 and 135,852 students with 817,988 and 1,026,096 term marks; 1,424 and 1,479 primary schools in both fits.*

The student-fixed-effect version's feeder effects are a different object by construction; they order feeders much as the baseline wedge does in 2022, less closely in 2018 (Table 62). The difference lies in the Grade 7 marks, not in the student effect as such or in the Grade 8 marks. The student effect averages a student's four Grade 7 term marks, and once the baseline's Grade 7 regressor is that same average the two order feeders alike (Spearman 0.967 in 2018 and 0.964 in 2022); averaging the four Grade 8 terms in place of Term 2 leaves a feeder's effect where it was (correlation 0.994 in 2018, net of sampling error). Each Grade 7 term's mark carries a part specific to the feeder and the term. Across feeders, the mean mark in one term departs from the same students' mean in the other three with a standard deviation of 0.39 to 0.46 within-province SD for Terms 1 to 3 in 2018 and 0.43 for Term 4. This part is not sampling noise: when each feeder's students are split at random into two halves, the halves agree on the Term 4 part with a reliability of 0.95. Nor is it learning that lasts, since feeders whose Term 4 marks run high against their earlier terms do not send students who do better in Grade 8 against those earlier marks (slope 0.04, standard error 0.06, where

learning would carry over at about 0.75). In 2022 the Term 4 mark is in effect a year mark built from the earlier terms, its own part shrinks to 0.16 SD, and the baseline and the student effect read nearly the same average. The baseline wedge is therefore the difficulty of the assessments behind a feeder’s Term 4 report mark, the mark that decides promotion; in 2018 that difficulty includes a part specific to the term.

References

- Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. “High Wage Workers and High Wage Firms.” *Econometrica* 67 (2): 251–333. <https://doi.org/10.1111/1468-0262.00020>.
- Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2026. “A Users’ Guide to Uncovering Worker and Firm Effects: The ABC of AKM.” *arXiv Preprint arXiv:2603.17034*, ahead of print. <https://doi.org/10.48550/arXiv.2603.17034>.
- Card, David, Ana Rute Cardoso, and Patrick Kline. 2016. “Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women.” *Quarterly Journal of Economics* 131 (2): 633–86. <https://doi.org/10.1093/qje/qjv038>.
- Card, David, Jörg Heining, and Patrick Kline. 2013. “Workplace Heterogeneity and the Rise of West German Wage Inequality.” *Quarterly Journal of Economics* 128 (3): 967–1015. <https://doi.org/10.1093/qje/qjt006>.
- Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” *American Economic Review* 104 (9): 2593–632. <https://doi.org/10.1257/aer.104.9.2593>.
- Dorans, Neil J., and Paul W. Holland. 2000. “Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case.” *Journal of Educational Measurement* 37 (4): 281–306. <https://doi.org/10.1111/j.1745-3984.2000.tb01088.x>.
- Fuller, Wayne A. 1987. *Measurement Error Models*. Wiley Series in Probability and Mathematical Statistics. Wiley. <https://doi.org/10.1002/9780470316665>.
- Holland, Paul W., and Neil J. Dorans. 2006. “Linking and Equating.” In *Educational Measurement*, 4th ed., edited by Robert L. Brennan. American Council on Education; Praeger.
- Hsieh, Chang-Tai, and Miguel Urquiola. 2006. “The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile’s Voucher Program.” *Journal of Public Economics* 90 (8–9): 1477–503. <https://doi.org/10.1016/j.jpubeco.2005.11.002>.
- Kane, Thomas J., and Douglas O. Staiger. 2008. *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. NBER Working Paper No. 14607. National Bureau of Economic Research. <https://doi.org/10.3386/w14607>.
- Kline, Patrick. 2024. “Firm Wage Effects.” Chap. 2 in *Handbook of Labor Economics*, vol. 5. Elsevier. <https://doi.org/10.1016/bs.heslab.2024.11.005>.
- Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2020. “Leave-Out Estimation of Variance Components.” *Econometrica* 88 (5): 1859–98. <https://doi.org/10.3982/ECTA16410>.
- Kolen, Michael J., and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*. 3rd ed. Springer. <https://doi.org/10.1007/978-1-4939-0317-7>.

- Kramarz, Francis, Stephen Machin, and Amine Ouazad. 2015. "Using Compulsory Mobility to Identify School Quality and Peer Effects." *Oxford Bulletin of Economics and Statistics* 77 (4): 566–87. <https://doi.org/10.1111/obes.12076>.
- Lachowska, Marta, Alexandre Mas, Raffaele Saggio, and Stephen A. Woodbury. 2023. "Do Firm Effects Drift? Evidence from Washington Administrative Data." *Journal of Econometrics* 233 (2): 375–95. <https://doi.org/10.1016/j.jeconom.2021.12.014>.
- Liu, Jing, Seth Gershenson, and Max Anthenelli. 2025. *Ready or Not? A New Way to Measure Elementary and Middle School Quality*. Report. Thomas B. Fordham Institute.
- Masters, Geoffrey N. 1988. "Anchor Tests, Score Equating and Sex Bias." *Australian Journal of Education* 32 (1): 25–43. <https://doi.org/10.1177/000494418803200102>.
- Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204. <https://doi.org/10.1080/07350015.2016.1227711>.
- Pop-Eleches, Cristian, and Miguel Urquiola. 2013. "Going to a Better School: Effects and Behavioral Responses." *American Economic Review* 103 (4): 1289–324. <https://doi.org/10.1257/aer.103.4.1289>.
- Postel-Vinay, Fabien, and Jean-Marc Robin. 2002. "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity." *Econometrica* 70 (6): 2295–350. <https://doi.org/10.1111/j.1468-0262.2002.00441.x>.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58. <https://doi.org/10.1111/j.1468-0262.2005.00584.x>.
- Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *The Quarterly Journal of Economics* 125 (1): 175–214. <https://doi.org/10.1162/qjec.2010.125.1.175>.
- Tartova, Desislava. 2025. "Teacher Value-Added in the Absence of Annual Test Scores: Utilizing Teacher Networks." Unpublished manuscript. February.