Online Appendix — Making School-Set Assessments Comparable: How School Assessments Hide Learning Inequality

Peter Courtney

November 11, 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 P for replication details.

Appendix A: Where each threat is settled

The appendices after this one are ordered by the threats to the paper's claims. Appendices B to E give the groundwork: the model and its assumptions, what the calibration claims, why the paper uses mathematics, and the six provinces. Each of Appendices F to O takes one threat, opens on its verdict and gives the checks behind it; Appendix P covers replication. Table 1 lists the ten threats, the checks that settle each and where they are; Appendix B.9 lists the assumptions the threats would break.

Table 1: **Each threat to the estimates, the checks that settle it, and where they are**

The estimates could be wrong if	Checks	Where
the wedge is estimation noise, or belongs to one cohort	Leave-out noise share; split halves; persistence across cohorts and a middle cohort	Appendix F
the wedge measures what a primary school left its students with, not how hard its assessments were	Learning during Grade 8; the Grade 8 anchor term; the order within groups of schools. Maintained: the holiday, and learning Grade 7 did not test	Appendix G
students choose secondary schools on something their grade missed	Grades before the move; a within-feeder permutation test; a student fixed effect; an endogenous-mobility bound	Appendix H
students who cannot be followed, or too few links between schools, distort the wedges	Linkage against the wedge; inverse-probability weighting; the connected set; the leave-out correction	Appendix I
a school does not grade all its students by one shift	Residuals by grade and wedge; a slope and shift per quintile; wedges by student group; other functional forms. Maintained: classes	Appendix J
grades pushed over the promotion threshold drive the wedge	De-heaping both grades; simulated heaping, uniform and graded by the school's mean grade	Appendix K
second-stage results overstate their precision	A bootstrap with every wedge redrawn; a Romano–Wolf adjustment	Appendix L

The estimates could be wrong if	Checks	Where
the common scale does not measure learning	Prediction of the NSC within secondary schools	Appendix M
the sorting result reflects co-movers, geography or ability	Leave-own-feeder-out; movers who cross a boundary; the other cohort's wedges; published matric	Appendix N
the value-added estimates record intake, peers or a school's own grading	A placebo on a grade earned before entry; moderation of school-based grades	Appendix O

Notes: Each row states a way the estimates could be wrong, the checks that address it, and the appendix that reports them. "Maintained" marks a condition the data cannot test. Table 4 of the main text summarises the standard checks of the two-way fixed-effects literature; Appendix A.1 says what each of its rows measures.

A.1: The two-way fixed-effects checks

Table 4 of the main text reports the standard checks of the two-way fixed-effects literature on the wedge, for the six provinces with Gauteng's values in brackets; this subsection says what each of its rows measures and on which sample, giving Gauteng's figures unless a row's sample below says otherwise.

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 grade. 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-grade grouping, the highest raw mean Grade 8 grades. The residual rows use pairs of deciles holding at least 100 students, by wedge decile × destination decile for the additivity row and by the student's Grade 7 decile × 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 grades with persistence held at one, so its school effects are net of each student's own effect rather than of the Grade 7 grade. 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 grades, cohorts pooled (165,408 students on split halves, 144,029 on the other cohort's estimates, 165,458 on mean grades); the coverage row every province and cohort of the National (6 provinces) sample; the sorting row 22,142 feeder cohorts 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 grades 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 samples; Gauteng not included).

Appendix B: The model and its assumptions

This appendix derives the estimating equation of §3.1 of the main text and states every condition it rests on. Three primitives, a Grade 7 measurement equation, a Grade 8 measurement equation and a law for how learning carries over, give four reduced forms, and with them the result that the slope on the Grade 7 grade converges to the persistence of learning times the reliability of the grade, $\lambda\rho$ (Appendix B.1 to Appendix B.8). Appendix B.9 lists the identifying assumptions in one place, Appendix B.10 shows that the two schools' effects add as a consequence of them, and Appendix B.11 translates the model into the two-way fixed-effects terms of the labour literature. The slope is 0.541 in 2018 and 0.573 in 2022, so persistence and reliability are each at least that large; Appendix B.12 to Appendix B.15 separate the two.

B.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^7_{ij}	latent (true) learning of student i at end of Grade 7, origin school j	unobserved
a^8_{ijk}	latent learning of student i at Grade 8 Term 2, destination cohort k	unobserved
g^7_{ij}	observed Grade 7 grade (Gr7Term4)	observed
g^8_{ijk}	observed Grade 8 grade (Gr8Term2)	observed
s_j	origin grade-scale offset: school-level shift between true learning and the Grade 7 grade	unobserved, structural
d_k	destination grade-scale offset: cohort-level shift between true learning and the Grade 8 grade	unobserved, structural
ε^7_{ij}, ε^8_{ijk}	idiosyncratic measurement error in each grade	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 grade (signal share of within-school variance)	structural, $\in (0, 1]$
β_1	slope on the lagged grade 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 grades below true learning:

$$g^7_{ij} = a^7_{ij} - s_j - \varepsilon^7_{ij}, \qquad \text{equivalently} \qquad a^7_{ij} = g^7_{ij} + s_j + \varepsilon^7_{ij}. \tag{B.1}$$

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 B.12.

B.2: The two measurement equations and the transition

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

$$a^8_{ijk} = g^8_{ijk} + d_k + \varepsilon^8_{ijk}, \qquad \text{equivalently} \qquad g^8_{ijk} = a^8_{ijk} - d_k - \varepsilon^8_{ijk}. \tag{B.2}$$

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

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

$$a^8_{ijk} = c + \lambda\, a^7_{ij} + \nu_{ijk}, \qquad 0 < \lambda \leq 1. \tag{B.3}$$

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. The main text lets the average gain differ by secondary school, c_k. Because it is common to every student in a destination cohort, it enters γ_k alongside $-d_k$ and leaves α_j unchanged, so this appendix writes a single c. The same fact keeps d_k from being identified on its own: a secondary school's gain and its difficulty enter only as their difference, $\gamma_k = c_k - d_k$.

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

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

Step A. Start from (B.2):

$$g^8_{ijk} = a^8_{ijk} - d_k.$$

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

$$g^8_{ijk} = a^7_{ij} + c + \nu_{ijk} - d_k.$$

Step C. Substitute (B.1) with $\varepsilon^7 = 0$, so $a^7_{ij} = g^7_{ij} + s_j$:

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

Step D. Collect by type:

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

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

$$\boxed{g^8_{ijk} = c + 1 \cdot g^7_{ij} + s_j - d_k + \varepsilon^*_{ijk}.} \tag{B.4}$$

Under (B.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.

B.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 (B.4) exactly, giving the same structural form, except the composite error now includes the measurement errors. The regressor $g^7_{ij} = a^7_{ij} - s_j - \varepsilon^7_{ij}$ 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 deviation net of both schools' effects:

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

The OLS probability limit is

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

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 $\mathrm{Cov}(\tilde{g}^7, \tilde{\varepsilon}^*)$ expands into nine terms:

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

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

$$\underbrace{\mathrm{Cov}(\tilde{a}^7, \tilde{\varepsilon}^7)}_{0} - \underbrace{\mathrm{Cov}(\tilde{a}^7, \tilde{\varepsilon}^8)}_{0} + \underbrace{\mathrm{Cov}(\tilde{a}^7, \tilde{\nu})}_{0} - \underbrace{\mathrm{Cov}(\tilde{\varepsilon}^7, \tilde{\varepsilon}^7)}_{\mathrm{Var}(\tilde{\varepsilon}^7)} + \underbrace{\mathrm{Cov}(\tilde{\varepsilon}^7, \tilde{\varepsilon}^8)}_{0} - \underbrace{\mathrm{Cov}(\tilde{\varepsilon}^7, \tilde{\nu})}_{0} = -\mathrm{Var}(\tilde{\varepsilon}^7).$$

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

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

Combining:

$$\hat{\beta}_1 \xrightarrow{p} \frac{\mathrm{Var}(\tilde{a}^7)}{\mathrm{Var}(\tilde{a}^7) + \mathrm{Var}(\tilde{\varepsilon}^7)} \equiv \rho. \tag{B.5}$$

The reliability ρ is the signal share of the within-school variance of the Grade 7 grade. 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.

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

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

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

$$g^8_{ijk} = a^8_{ijk} - d_k.$$

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

$$g^8_{ijk} = (c + \lambda a^7_{ij} + \nu_{ijk}) - d_k.$$

Step C. Substitute (B.1) with $\varepsilon^7 = 0$, so $a^7_{ij} = g^7_{ij} + s_j$:

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

Step D. Distribute λ and collect:

$$\boxed{g^8_{ijk} = c + \lambda\, g^7_{ij} + \lambda s_j - d_k + \nu_{ijk}.} \tag{B.6}$$

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

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

since $\mathrm{Cov}(\tilde{g}^7, \tilde{\nu}) = 0$ 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.

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

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

Step A. Start from (B.2):

$$g^8_{ijk} = a^8_{ijk} - d_k - \varepsilon^8_{ijk}.$$

Step B. Substitute (B.3):

$$g^8_{ijk} = (c + \lambda a^7_{ij} + \nu_{ijk}) - d_k - \varepsilon^8_{ijk}.$$

Step C. Substitute (B.1), $a^7_{ij} = g^7_{ij} + s_j + \varepsilon^7_{ij}$:

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

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

$$\boxed{g^8_{ijk} = c + \lambda\, g^7_{ij} + \lambda s_j - d_k + \varepsilon^{**}_{ijk}.} \tag{B.7}$$

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, $\tilde{g}^7 = \tilde{a}^7 - \tilde{\varepsilon}^7$:

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

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

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

$$\boxed{\hat{\beta}_1 \xrightarrow{p} \lambda\rho.} \tag{B.8}$$

The estimated slope is persistence times reliability. From $\hat{\beta}_1 \approx 0.54$ alone, λ and ρ are not separately identified: any pair with $\lambda\rho = 0.54$ fits the data equally. However, since both factors lie in $(0, 1]$, each is bounded below by the product: $\lambda \geq 0.54$ and $\rho \geq 0.54$. This one-sided bound requires no additional assumptions beyond the model above.

B.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:

$$\hat{\alpha}_j \approx \underbrace{\overline{g^8}_j}_{\text{cohort anchor performance}} - \beta_1 \underbrace{\overline{g^7}_j}_{\text{cohort raw grade}} . \tag{B.9}$$

The three reference points are:

- **At** $\beta_1 = 0$ (no lagged control): $\hat{\alpha}_j = \overline{g^8}_j$, 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** $\beta_1 \approx 0.54$ (what the data yield): $\hat{\alpha}_j$ is a blend: anchor performance net of a 0.5-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 grades) receives the higher $\hat{\alpha}_j$. This is an algebraic property of the estimator, not a causal production claim. Under equation (B.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 13–14% of the no-fee contrast, less than the shrinkage hides (Appendix G.5).

The strong negative correlation between $\hat{\alpha}_j$ and raw Grade 7 means within each province and cohort (approximately -0.58) is partly mechanical: equation (B.9) subtracts $\beta_1 \overline{g^7}_j$ 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 × destination) estimator is fixed by the full system of normal equations across the connected mobility network, not literally by $\overline{g^8}_j - \beta_1 \overline{g^7}_j$ school by school. Equation (B.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.

B.8: Technical summary

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

The headline estimated slope $\hat{\beta}_1 \approx 0.54$ sits in the last row. Decomposing λ from ρ requires extra information; Appendix B.12 uses the English grade 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 (B.8), and the $\hat{\alpha}_j$ continuum are summarised for the main text in §3.1 of the main text.

B.9: The complete set of identifying assumptions

Sections B.1–B.2 fixed the structural primitives and §B.4–B.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: $a^7_{ij} = g^7_{ij} + s_j + \varepsilon^7_{ij}$, with no school-specific scale stretching ($\theta_j \equiv 1$) and s_j independent of a student's position in the grade distribution (B.1).	Gives a single slope and a scalar α_j per school. A school-specific θ_j would appear as an origin-specific slope (§B.10). Slopes do differ by origin, through within-school reliability rather than θ_j, and the order of schools at the average grade barely moves (Appendix J.1).
M2	Destination offset d_k is additive, common to all students in destination cohort k, and independent of origin j (B.2).	Makes γ_k a single destination term, $c_k - d_k$: the cohort's early gain net of its difficulty, not difficulty alone. This is the main text's *common within-grade anchor*; warranted by one assessment per destination cohort 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 (§B.1).	Fixes the slope's attenuation to the scalar ρ (B.5, B.8). Relaxations are examined in Appendix B.12.
D1	Homogeneous linear persistence: a single $\lambda \in (0, 1]$, constant across students, origins and destinations (B.3).	*Generates* additive separability (§B.10) and the single slope. A destination-specific λ_k breaks separability; the slope does differ across destinations, but fitting it barely reorders feeders (§B.10).
D2	Growth-shock exogeneity: $\mathrm{E}[\nu_{ijk} \mid a^7_{ij}, \varepsilon^7, \varepsilon^8] = 0$.	With M3, delivers the probability-limit algebra of §B.6.
X1	Exogenous mobility: conditional on (g^7_{ij}, 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 H.2, Appendix H.1.

Axiom	Statement	Role, and where disciplined
X2	Connectedness: within each province and cohort the bipartite origin × destination graph has a giant connected component; relative α_j are identified only within it.	Chains the within-grade comparisons into one scale. Coverage: main text and Appendix I.1.
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 I.
R1	Large feeders, dense network: each origin links many students to many destinations, so the origin-mean measurement error $\bar{\varepsilon}_j^7 \to 0$ 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 Gauteng, the highest of the six provinces (Table 6); limited-mobility bias in Appendix I.3.

In these terms the main text's identifying conditions are exogenous mobility (X1), within a connected set (X2); additivity, which §B.10 derives from M1, M2 and D1; the exclusion restriction, which no axiom above states; and, among the conditions specific to school-set grades, the common within-grade anchor (M2) and class grading unrelated to the feeder. M3 and D2 are the classical conditions of §B.1 that the plim algebra consumes; S1 and R1 are regularity conditions. Sections B.3–B.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 $\hat{\alpha}_j$.

B.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 §B.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 grade decomposes as

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

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

Proof. This is the general reduced form (B.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 $c_k - d_k$, which by M2 and the transition law has 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 $\lambda\rho$ (B.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 (§B.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 J.1).

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 §B.6 instead yields

$$g^8_{ijk} = c + \lambda_k\, g^7_{ij} + \lambda_k s_j - d_k + \varepsilon^{**}_{ijk},$$

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

(ii) *Scale stretching.* If difficulty loads non-unitarily, $a^7_{ij} = \theta_j g^7_{ij} + s_j + \varepsilon^7_{ij}$ with $\theta_j \neq 1$, the slope becomes origin-specific ($\lambda\theta_j$) and $\hat{\alpha}_j$ depends on the shape of a school's grade 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 grade on the Grade 7 grade 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 grades 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 grade, 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 mean residual 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 J.1). 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 student's grade. A slope per feeder takes up most of that variation and leaves the order of feeders at the average grade 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 Grade 7 and Grade 8 grades, the signature of a higher within-school reliability ρ_j rather than a different scale (§3.3 of the main text; Appendix J.1).

B.11: 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 assessment: the feeder in Grade 7, the destination cohort in Grade 8 (Table 2).

Table 2: **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 assessment: feeder j in Grade 7, destination cohort k in Grade 8
Log wage	The grade, standardised within province and cohort
Worker effect	Latent learning at the end of Grade 7, a^7_{ij}
Firm effect	The school's grade-scale offset, s_j for the feeder and d_k for the destination, estimated as $\alpha_j = \lambda s_j$ and $\gamma_k = c_k - d_k$ (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

Two-way fixed-effects model (labour)	This paper
The one departure	Learning need not carry over one for one: the slope β_1 on the Grade 7 grade is left free, and it estimates $\lambda\rho$, persistence times the Grade 7 grade's reliability (Appendix B.12)

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 H.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 grades of students bound for different destinations ran nearly parallel before it (Appendix H.1). And each student moves once, so a grade 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 grade: 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 grades decide promotion, so the estimation sample is selected on the grades being calibrated (§2 of the main text). At a combined school the feeder and the destination cohort belong to one institution, so its stayers compare the school's Grade 7 grading with its own Grade 8 grading.

B.12: The slope as persistence times reliability

The latent-variable derivation in Appendix B.1 to Appendix B.8 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-school reliability (signal share) of the Grade 7 grade. The OLS estimate, $\hat{\beta}_1 = 0.541$ (2018 cohort) and 0.573 (2022 cohort), is therefore a product of two distinct factors that the baseline regression cannot separate. The purpose of these four subsections is to decompose $\hat{\beta}_1$ into its two components using two independent tests, and to reconcile the tests into a single set of bounds on λ and ρ.

One-sided bounds without any test. From $\lambda\rho \approx 0.54$ and both factors bounded in $(0, 1]$, each factor satisfies $\lambda \geq 0.54$ and $\rho \geq 0.54$ 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 54% persistent and the Grade 7 grade is at least 54% reliable signal. The two tests below sharpen these bounds into a narrower interval.

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

1. **Test 1 (IV).** Instrument the Grade 7 Mathematics grade with the student's Grade 7 English FAL grade. 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-school reliability ρ of the Term 4 grade from pairs of the same student's within-school-demeaned Grade 7 term grades that share no component. The Term 1 and Term 4 grades are not such a pair: in 2022 the Term 4 grade is in effect a year grade built from the earlier terms' grades, and in both cohorts shared heaping at the 40% threshold (the count at $(T_1, T_4) = (40, 40)$ is 6.09× its expected frequency) adds to the overlap. Terms 1, 2 and 3 do not contain one another.

B.13: Test 1: IV decomposition

The instrumental-variables estimator uses the Grade 7 English FAL Term 4 grade as an instrument for the Grade 7 Mathematics Term 4 grade, within the same origin-school × destination-cohort fixed-effects structure as the headline OLS regression. The identifying assumption is that the measurement errors in the two subjects' grades are uncorrelated conditional on the school FE: $\mathrm{Cov}(\tilde{\varepsilon}^7_{\mathrm{Eng}}, \tilde{\varepsilon}^7_{\mathrm{Maths}}) = 0$ after within-school demeaning. The student-level cross-subject heaping correlation at 40% is $r \approx 0.05$–0.10, substantially weaker than the within-subject cross-term correlation of $r = 0.27$ that disqualified same-subject earlier-term grades. 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 3: 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 grade for the same student. Scale: both grades are z-scored within province × cohort × subject, as in the headline regression, so both slopes are in SD of the Grade 8 grade per SD of the Grade 7 grade, and implied ρ is the ratio of two slopes fitted on the same students. Identifying assumption: $\mathrm{Cov}(\tilde{\varepsilon}^7_{Eng}, \tilde{\varepsilon}^7_{Maths}) = 0$ *after within-school 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 ∩ English FAL student IDs.*

With both grades 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 headline slopes of 0.541 and 0.573 because the sample keeps only students with an English FAL grade; in Gauteng those students give 0.580 and 0.576, against 0.601 and 0.592 for all its students. First-stage F-statistics range from 17,098 to 105,118 across the 12 province-cohort samples, 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 B.15). 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 grade, 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).

B.14: Test 2: Direct reliability estimation

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

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

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 ρ^{KS}, 0.84 in 2018 and 0.87 in 2022 across the six provinces, 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 grade variance is true learning signal, a different question and a different quantity.

The test–retest estimator correlates two within-school-demeaned grades 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 4 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 grade is in effect a year grade: within a school the Term 1 to Term 3 grades 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 grade sharing no error with the others (0.83 in 2018). Both pairs also share heaping: the count at $(T_1, T_4) = (40, 40)$ is 6.09× its expected frequency under independence. The restated test uses only the three pairs among Terms 1 to 3.

Table 4: 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 Gauteng's OLS slope. IV columns reproduced for each province and cohort from Table 3 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
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) grade against Grade 7 Term 4 grade, for the same student. Schools with fewer than 5 students dropped. Implied $\hat{\lambda} = \hat{\beta}_1^{OLS}/\hat{\rho}$, using Gauteng's $\hat{\beta}_1^{OLS}$, 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 grade is built largely from the earlier terms' grades, 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 values for the same province and cohort from Table 3, on one scale. Sample: National six-province Mathematics panel, intersection of Grade 7 T1 (or T3) and T4 grade 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. The Term 1 with Term 4 estimate exceeds the IV-implied ρ in 9 of the 12 province-cohort samples of Table 4, and the Term 3 with Term 4 estimate in 11. Every exception is a 2018 sample: 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 grade 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 grade the earlier terms do not predict does not load negatively on learning. One term's grade 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 grade is built explains the rest of the rise, not heaping. The Term 1 to Term 3 grades explain 71% of the Term 4 grade'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 grade that is part of the Term 4 grade does not retest it.

B.15: Synthesis: bounds on λ and ρ

The two 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 grade retests the Term 4 grade 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).

Combining the IV and test–retest bounds:

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

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 grade'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 grade. Taking the test–retest floor as the conservative bound:

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

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 $r \approx 0.05$–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 grade's stable share within the year, which no Grade 7 retest can pin down more tightly once the Term 4 grade contains the earlier terms. The two are not computed on the same students. The test–retest λ divides Gauteng's slope by a six-province reliability; the IV slope and its ρ come from the six-province students with an English FAL grade, whose OLS slope is lower (0.522 against Gauteng's 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 grade carries at most about 70% (2018) and 77% (2022) true learning signal within a school. Genuine persistence fade ($\lambda < 1$, Story 2) is present in both cohorts: 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 $\lambda \geq 0.54$.

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 et al. (2011), who address attenuation explicitly with dynamic-panel methods on Pakistani data, find that "only a fifth to a half of learning persists between grades", an attenuation-corrected persistence of roughly 0.2–0.5 over annual horizons;[1] the gap between their figures and ours is consistent with a shorter horizon and a different institutional setting.

Appendix C: Moderation, not equating

Psychometrics distinguishes several ways of putting scores from different tests on one scale, and each warrants a different claim. The calibration is statistical moderation, or scale aligning: the shared Grade 8 assessments are the anchor, and a school's wedge is the constant that places its grades on the anchor's scale (Appendix C.5). It is not equating, because Grade 7 assessments are not built to common specifications and no student's grades are interchangeable across schools (Appendix C.3), and it is not projection, because it does not predict Grade 8 grades from Grade 7 ones. Moderation is sound only if the anchor measures what the school's assessments measure (Appendix C.4), the condition Appendix G.3 tests. Appendix C.1 and Appendix C.2 set out the two classifications the placement draws on, Mislevy and Linn's and Holland and Dorans's.

C.1: 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 grades 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.

C.2: 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 grades on the scale defined by the Grade 8 within-grade anchor, rather than predicting or equating.

[1] In 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 et al. (2017), which concerns an information intervention rather than persistence estimation.

C.3: 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.

C.4: 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 G.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.

C.5: 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 grades; 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 grade times the Grade 7 slope $\hat{\beta}_1$ plus her school's $\hat{\alpha}_j$, and that sum places each school's grades 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 D: Why mathematics only

The paper estimates wedges from mathematics grades alone. A school's mathematics and English are taught by different teachers and assessed separately, so one model for both would force one slope on two subjects and give a school effect that blends them (Appendix D.1). Estimated on its own, the English First Additional Language calibration is too noisy to carry as a measure (Appendix D.2), although a school's English and mathematics wedges move together: disattenuated for estimation error, they correlate at 0.601 (§4.2 of the main text).

D.1: 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.

D.2: Three-way comparison of key statistics

Table 5: 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 (points per quintile step)	1.93	0.73	1.92

Sample: Mathematics and English columns are for Gauteng, pooled 2018 and 2022 cohorts. The Pooled column, both subjects in one model, comes from an earlier run on 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.

Appendix E: The six provinces

Each province is estimated on its own, on its own connected set, and the main text reports the six together, naming Gauteng where a result is Gauteng's alone. Appendix E.1 gives each province and cohort's diagnostics: the reliability of the wedges is 0.93 or more in every one, highest in Gauteng and lowest in the Eastern Cape. Appendix E.2 applies the same specification, unchanged, to a separate national extract of the same records, 31 estimates across 8 provinces and 3 cohorts, where the slope runs from 0.475 to 0.770 and Gauteng's slope and spread, from the paper's own extract, lie inside both ranges. Appendix E.3 gives every number the main text cites, across the six provinces and in each province.

E.1: Province diagnostics and robustness

Per-province estimation diagnostics for both cohort years. Each row reports, for that province and cohort, 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 province-cohort 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 6: 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

(continued)

Province	Cohort	Schools	Students	β_1	Median SE	Reliability	SNR	Verdict
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.

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

E.2: 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, 31 times in Mathematics, each for a province, a Grade 7 cohort and a Grade 8 anchor term (Table 7, Panel A). The estimates span 3 cohorts, those of 2021 to 2023, and 8 provinces. The Grade 8 anchor in these estimates is the Term 1 or the Term 4 grade; the paper's own anchor is the Term 2 grade, which the national extract does not carry. Gauteng is not among those provinces: the national extract holds no Grade 7 Term 4 grade for Gauteng in these cohorts, so the province has no estimate. Panel B reproduces the paper's own estimates for Gauteng 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 estimates of Panel A the slope of the standardised Grade 8 anchor grade on the standardised Grade 7 grade 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 grade. The lowest slope and the highest dispersion both come from a single thin sample, North West's 2022 cohort at the Term 1 anchor, which rests on few feeder schools. Gauteng's slope and dispersion in Panel B lie inside both ranges.

Table 7: **The estimator returns the same slope and dispersion in 31 out-of-sample replications.**

	Slope (1)	SD (2)	Med. SE (3)	SNR (4)	Sig. (%) (5)	R2 (6)	Learners (7)	Schools (8)
Panel A. This extract, national, Mathematics								
EC 2021 T1	0.678	0.627	0.091	6.895	66.7	0.503	13,239	6
	(0.042)							
EC 2021 T4	0.721	0.700	0.078	8.925	100.0	0.567	11,760	6
	(0.044)							
EC 2022 T4	0.705	0.835	0.096	8.694	78.2	0.513	22,142	435
	(0.006)							
EC 2023 T1	0.644	0.847	0.105	8.077	79.0	0.417	21,276	428

(continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	(0.006)							
EC 2023 T4	0.737	0.896	0.092	9.720	77.8	0.552	21,695	451
	(0.006)							
FS 2023 T1	0.718	0.679	0.090	7.536	79.5	0.472	5,782	78
	(0.013)							
FS 2023 T4	0.770	0.744	0.084	8.823	88.0	0.571	5,565	83
	(0.012)							
KZ 2021 T1	0.550	0.450	0.101	4.458	63.1	0.333	179,591	3,394
	(0.002)							
KZ 2021 T4	0.670	0.505	0.093	5.430	68.5	0.466	180,005	3,401
	(0.002)							
KZ 2022 T1	0.594	0.460	0.101	4.577	64.7	0.372	184,627	3,459
	(0.002)							
LP 2021 T1	0.529	0.439	0.100	4.398	63.4	0.324	112,221	2,228
	(0.003)							
LP 2021 T4	0.630	0.470	0.086	5.487	69.7	0.473	113,393	2,243
	(0.002)							
LP 2022 T1	0.554	0.455	0.098	4.659	65.0	0.336	113,952	2,246
	(0.003)							
LP 2022 T4	0.669	0.517	0.087	5.949	72.3	0.493	76,877	1,583
	(0.003)							
LP 2023 T1	0.544	0.522	0.084	6.176	72.5	0.374	10,648	207
	(0.008)							
LP 2023 T4	0.705	0.591	0.071	8.311	82.3	0.585	10,750	204
	(0.007)							
MP 2021 T1	0.637	0.681	0.087	7.870	79.0	0.411	16,835	252
	(0.007)							
MP 2021 T4	0.720	0.682	0.075	9.071	77.5	0.534	16,815	249
	(0.006)							
MP 2022 T1	0.596	0.502	0.086	5.828	69.2	0.364	69,632	1,017
	(0.003)							
NC 2021 T1	0.546	0.598	0.083	7.230	76.0	0.397	18,371	300
	(0.006)							
NC 2021 T4	0.654	0.605	0.076	7.957	77.9	0.531	17,950	299
	(0.005)							
NC 2022 T1	0.609	0.572	0.085	6.735	72.9	0.415	19,480	306
	(0.006)							
NW 2021 T1 †	0.579	0.720	0.108	6.660	87.5	0.396	2,918	24
	(0.027)							
NW 2021 T4 †	0.579	0.721	0.108	6.664	87.5	0.396	2,920	24
	(0.027)							
NW 2022 T1 †	0.475	1.271	0.100	12.747	76.5	0.274	2,878	51
	(0.019)							
NW 2022 T4 †	0.588	1.118	0.095	11.815	77.5	0.401	2,714	49
	(0.020)							
NW 2023 T1	0.713	0.668	0.081	8.202	86.0	0.593	3,091	50
	(0.013)							
NW 2023 T4	0.761	0.654	0.072	9.091	76.5	0.669	3,126	51
	(0.012)							
WC 2022 T1	0.482	0.589	0.189	3.118	55.1	0.202	1,008	49
	(0.040)							

(continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
WC 2022 T4	0.575	0.701	0.177	3.960	54.2	0.300	1,003	48
	(0.039)							
WC 2023 T4	0.736	0.507	0.064	7.944	77.6	0.593	54,417	890
	(0.003)							
Panel B. Gauteng core, Grade 8 Term 2 anchor								
GT 2018 T2	0.601	0.471	0.067	7.000	70.4	0.404	106,514	1,424
	(0.003)							
GT 2022 T2	0.592	0.468	0.067	6.990	74.8	0.398	135,852	1,479
	(0.002)							

Note. Column (1), Slope, is the slope of the Grade 8 anchor grade on the Grade 7 baseline grade, with its standard error in parentheses on the line beneath. Columns (2) to (5) describe the estimated intake-school effect in each row: 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 that row's within R-squared; columns (7) and (8) are its learners and its Grade 7 feeder schools. Each row of Panel A is one province, cohort and anchor term, estimated with the paper's own specification unchanged: a valid grade at both ends, a minimum origin-school size of ten, grades standardised within province by cohort by subject, the largest connected component, and samples of fewer than 100 students skipped. The row label gives the province, the Grade 7 cohort year and the Grade 8 anchor term. Effects are in within-province standard deviations of the standardised grade. Panel A holds the 31 samples that estimate, out of the 54 Mathematics samples the extract defines: 20 are structurally absent because the grades of one of the two terms are not in the extract at all, and 3 are too thin to estimate. A dagger marks a row whose two anchor terms hold the same grades, so its anchor comparison is not informative; 4 rows carry it. Levels are not comparable between the two panels: the paper's Gauteng 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 sample-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 lists where each number comes from. **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 samples carry 1,003 to 184,627 learners and 6 to 3,459 feeder schools each. Panel B, the paper's own Gauteng Mathematics samples at the Grade 8 Term 2 anchor, 2018 and 2022 cohorts, 1,424 to 1,479 feeder schools.

E.3: Every number the main text cites, by province

Table 8 gives every number the main text cites, across the six provinces and in each province.

Table 8: **Every number the main text cites, across the six provinces and in each province.**

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
1	Students in the 2018 estimation sample (abstract and §1 print 2018+2022 in millions) (6P, 2018)	532,295	106,514	108,765	120,760	89,616	63,602	43,038

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
2	Students in the 2022 estimation sample (abstract and §1 print 2018+2022 in millions) (6P, 2022)	665,019	135,852	117,881	161,960	116,910	75,922	56,494
3	Precision-weighted correlation of school effects anchored on Grade 8 Term 1 vs Term 4 (headline pool) (DBE extract 2021–24, its provinces pooled)	0.902	—	0.763	0.96	0.942	0.84	0.834
4	Mean number of primaries feeding a secondary school, Mathematics (6P, 2018)	11.993	21.195	10.452	11.634	9.61	9.848	11.664
5	Lowest share of primaries inside the largest connected set (%) (6P, min over 12 province-cohort fits)	98.8	99.4	98.8 / 99.6	99.7	99.6	99.4	99.8
6	Split-half correlation of the wedge (6P, 2022, mean of 6 fits)	0.743	0.879	0.666	0.743	0.749	0.758	0.748
7	Precision-weighted 2018–2022 correlation of the wedge (6P, schools fitted in both cohorts)	0.477	0.624	0.457	0.478	0.365	0.409	0.456
8	Quintile 5 minus Quintile 1 gap in raw Grade 7 school means (SD), no controls (6P, 2018)	0.476	0.395	0.934	0.566	0.538	0.269	—
9	Quintile 5 minus Quintile 1 gap in raw Grade 7 school means (SD), no controls (6P, 2022)	0.234	0.174	0.545	0.259	0.445	0.041	—
10	Quintile 5 minus Quintile 1 gap in the calibrated level (SD), no controls (6P, 2018)	0.722	0.604	1.14	0.668	0.858	0.485	—
11	Quintile 5 minus Quintile 1 gap in the calibrated level (SD), no controls (6P, 2022)	0.766	0.707	1.182	0.683	0.871	0.473	—
12	Share of school-cohorts whose quintile of the calibrated level differs from their raw-mean quintile (%) (6P, 2018+2022 school-cohorts)	77.1	72.5	77.5	78	78.7	77.3	76.4
13	No-fee minus fee-charging wedge contrast (SD), precision-weighted (6P, 2018+2022 pooled)	0.413	0.47	0.496	0.216	0.51	0.417	0.519
14	Wedge coefficient on destination peers' Grade 8 mean, own pupil left out (6P movers, 2018+2022)	0.117	0.165	0.066	0.065	0.114	0.144	0.138
15	Raw Grade 7 school mean on peers' Grade 8 mean, own feeder left out (6P movers, 2018+2022)	0.012	0.020	0.021	0.011	-0.019	-0.008	0.003
16	p-value printed as stars on the row above (6P movers, 2018+2022)	0.084	0.173	0.148	0.394	0.234	0.694	0.882
17	Split-half reliability of exit-school effects theta, NSC Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.855	0.831	0.872	0.863	0.803	0.899	0.793
18	Conventional reliability benchmark (0.80), set in a §1 chunk (literature)	—	—	—	—	—	—	—
19	Share of estimation-sample students changing school, Grade 7 to 8 (%) (6P, 2018)	86.397	92.065	71.856	92.917	95.461	71.207	94.333
20	Share of estimation-sample students changing school, Grade 7 to 8 (%) (6P, 2022)	91.701	94.971	81.702	92.94	96.715	87.357	96.593
21	Median number of Grade 8 destinations per primary (6P, 2018)	6	12	4	6	5	6	5
22	Mean number of Grade 8 destinations per primary (6P, 2018)	7.677	13.454	5.895	8.459	6.384	6.802	6.378

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
23	Median number of primaries feeding a secondary school (6P, 2018)	8	20	6	8	6	8	9
24	Median Herfindahl index of a primary's leavers across destinations (text: both cohorts) (GT, 2018)	0.48	0.256	0.583	0.444	0.501	0.606	0.522
25	Lowest median leavers' Herfindahl index among the other five provinces (5 provinces, both cohorts (min))	0.444	0.256 / 0.259	0.521	0.444	0.501	0.517	0.513
26	Highest median leavers' Herfindahl index among the other five provinces (5 provinces, both cohorts (max))	0.606	0.256 / 0.259	0.583	0.446	0.51	0.606	0.522
27	Share of linked students at an independent primary or secondary (%) (6P, 2018)	3.8	7.307	4.592	1.293	4.418	1.403	2.362
28	Share of linked students at an independent primary or secondary (%) (6P, 2022)	3.7	6.064	4.641	1.409	4.815	2.213	2.155
29	Lowest rank correlation of the wedge refitted without independent schools with the full-sample wedge (6P, min over 12 fits)	0.994	0.997	0.997	0.999	0.994	0.997	0.999
30	Largest change in the Grade 7 slope without independent schools (6P, max over 12 fits)	0.012	0.001	0.004	0.002	0.012	<0.001	<0.001
31	SD of the wedge without independent schools over its full-sample SD (GT, 2018)	0.981	0.92	0.997	0.99	0.975	1.008	0.983
32	Share of Term 4 grade variance explained within school by Terms 1–3 (%) (6P, 2022)	94	95.304	92.453	93.032	93.974	93.334	94.368
33	Share of Term 4 grade variance explained within school by Terms 1–3 (%) (6P, 2018)	71	76.325	65.98	71.553	70.362	71.041	70.162
34	Students entering a secondary school with a calibrated intake score (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	394,353	45,084	85,425	86,185	85,411	54,432	37,816
35	Grade 7 enrolments in SA-SAMS (6P, 2018)	796,928	—	—	—	—	—	—
36	Grade 7 enrolments in SA-SAMS (6P, 2022)	866,679	—	—	—	—	—	—
37	Share of Grade 7 enrolments with a valid Term 4 mathematics grade (%) (6P, 2018)	88.9	—	—	—	—	—	—
38	Share of Grade 7 enrolments with a valid Term 4 mathematics grade (%) (6P, 2022)	92.8	—	—	—	—	—	—
39	Share of valid Grade 7 records linked to a Grade 8 record with a valid Term 2 grade (%) (6P, 2018)	76.8	67.594	80.964	80.62	76.925	81.059	75.427
40	Share of valid Grade 7 records linked to a Grade 8 record with a valid Term 2 grade (%) (6P, 2022)	83.8	74.385	84.998	86.543	88.768	89.186	82.515
41	Students in the estimation sample (6P, 2018)	532,295	106,514	108,606	120,718	89,601	63,557	43,038
42	Students in the estimation sample (6P, 2022)	665,019	135,852	117,770	161,960	116,910	75,922	56,439
43	SD of Grade 7 grades, in points (size of one SD) (6P, 2018)	13.867	15.342	11.743	13.911	13.073	14.326	14.646
44	SD of Grade 8 grades, in points (size of one SD) (6P, 2018)	17.716	19.438	17.1	17.738	16.223	17.165	17.127

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
45	Primaries outside the largest connected set of their province and cohort (6P, both cohorts)	103	13	51	15	12	9	3
46	Those primaries as a share of all primaries (%) (6P, both cohorts)	0.448	0.446	0.815	0.267	0.273	0.446	0.169
47	Share of students at those primaries (%) (6P, both cohorts)	0.171	0.104	0.445	0.114	0.087	0.14	0.086
48	School-years in the linkage check (6P, 12 fits pooled)	22,868	2,903	6,208	5,597	4,379	2,011	1,770
49	Correlation of a school's wedge with its linkage rate (6P, 12 fits pooled)	-0.034	0.091	-0.011	-0.096	-0.109	-0.02	-0.049
50	Shift in the mean wedge under inverse-linkage weighting (SD) (6P, 12 fits pooled)	0.004	-0.005	<0.001	0.01	0.008	0.002	0.008
51	Slope of the Grade 8 grade on the Grade 7 grade (beta1) (GT, 2018)	0.541	0.601	0.5	0.53	0.533	0.585	0.593
52	Slope of the Grade 8 grade on the Grade 7 grade (beta1) (GT, 2022)	0.573	0.592	0.526	0.585	0.58	0.607	0.61
53	Persistence: Grade 7 slope instrumented with the Grade 7 English FAL grade (6P, 2018)	0.746	0.76	0.74	0.712	0.721	0.771	0.838
54	Persistence: Grade 7 slope instrumented with the Grade 7 English FAL grade (6P, 2022)	0.737	0.734	0.669	0.762	0.732	0.759	0.793
55	Reliability of the Grade 7 grade (OLS slope over IV slope) (6P, 2018)	0.699	0.763	0.628	0.686	0.72	0.738	0.694
56	Reliability of the Grade 7 grade (OLS slope over IV slope) (6P, 2022)	0.769	0.784	0.758	0.752	0.784	0.783	0.759
57	SD of the wedge across primaries (SD of Grade 8 grades) (6P, 2018)	0.453	0.471	0.467	0.426	0.45	0.439	0.475
58	SD of the wedge recomputed at the IV slope (6P, 2018)	0.549	0.554	0.578	0.513	0.539	0.526	0.585
59	SD of the wedge across primaries (SD of Grade 8 grades) (6P, 2022)	0.453	0.468	0.446	0.44	0.449	0.465	0.486
60	SD of the wedge recomputed at the IV slope (6P, 2022)	0.527	0.527	0.51	0.527	0.523	0.543	0.573
61	No-fee minus fee wedge gap, OLS wedges (Table 4 sample, SD) (6P, 2018+2022 pooled)	0.413	0.47	0.496	0.216	0.51	0.417	0.519
62	No-fee minus fee wedge gap at the IV slope (SD) (6P, 2018+2022 pooled)	0.465	0.509	0.551	0.262	0.553	0.496	0.615
63	Correlation of OLS wedges with IV-slope wedges (6P, 2018)	0.986	0.994	0.979	0.989	0.987	0.988	0.983
64	Correlation of OLS wedges with IV-slope wedges (6P, 2022)	0.992	0.995	0.991	0.992	0.992	0.994	0.991
65	Intake's share of wedge variance, low end (%) (6P, pooled, range over named slopes)	10	9	12	9	10	8	10
66	Intake's share of wedge variance, high end (%) (6P, pooled, range over named slopes)	12	10	17	11	11	9	13
67	Intake's share of the no-fee gap, low end (%) (6P, pooled, range over named slopes)	13	13	15	7	16	6	11
68	Intake's share of the no-fee gap, high end (%) (6P, pooled, range over named slopes)	14	15	21	9	17	7	13
69	Share of linked students crossing a provincial border (%) (6P, both cohorts, both subjects)	1.876	3.269	1.596	0.868	1.908	1.626	2.7
70	Students per cross-province feeder–destination pair (6P, Mathematics, both cohorts)	1.142	1.116	1.078	1.075	1.12	1.178	1.446

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
71	Students per within-province feeder–destination pair (6P, Mathematics, both cohorts)	8.155	7.986	7.063	6.676	9.191	13.396	11.367
72	Number of provinces in the sample ("six") (6P)	—	—	—	—	—	—	—
73	Lowest noise-corrected correlation of other-term wedges with Term 2 wedges (6P, both cohorts)	0.964	0.985	0.941	0.96	0.961	0.959	0.956
74	Two-way clustered standard error of beta1 (6P, 2018, mean of 6 fits)	0.004	0.01	0.009	0.008	0.009	0.011	0.014
75	Two-way clustered standard error of beta1 (6P, 2022, mean of 6 fits)	0.003	0.007	0.007	0.006	0.007	0.01	0.011
76	P90 minus P10 of the wedge, averaged over fits (SD) (6P, mean over 12 province-cohort fits)	1.13	1.137	1.117	1.087	1.109	1.098	1.204
77	SD of the destination effects (SD) (6P, 2018, mean of 6 fits)	0.654	0.578	0.7	0.676	0.653	0.647	0.55
78	SD of the destination effects (SD) (6P, 2022, mean of 6 fits)	0.598	0.518	0.67	0.61	0.579	0.612	0.511
79	Share of the sample on stayer links, at combined schools (%) (6P, 2018)	13.603	7.935	28.144	7.083	4.539	28.793	5.667
80	Share of the sample on stayer links, at combined schools (%) (6P, 2022)	8.299	5.029	18.298	7.06	3.285	12.643	3.407
81	Grade 8 Term 2→4 drift per SD of the primary's wedge (6P, 2018+2022 pooled)	0.026	0.033	0.025	0.036	0.022	<0.001	0.025
82	Drift that proportional grading alone predicts (6P, 2018+2022 pooled)	0.037	0.042	0.04	0.045	0.034	0.013	0.036
83	Pre-move Grade 7 T1→T4 gain, top minus bottom destination quartile (split-half wedges) (6P, 2018+2022 pooled)	0.012	<0.001	0.009	0.015	0.019	0.022	0.008
84	Pre-move Grade 7 T1→T4 gain, top minus bottom destination quartile (other-cohort wedges) (6P, 2018+2022 pooled)	0.012	0.012	0.012	0.007	0.012	0.017	0.017
85	Estimation-noise share of wedge variance, leave-out correction (%) (6P, 2018)	16.051	5.394	20.776	15.648	17.425	18.331	13.881
86	Estimation-noise share of wedge variance, leave-out correction (%) (6P, 2022)	13.438	5.451	18.262	14.118	12.812	12.569	12.005
87	Largest mean residual net of the Grade 7 margin (additivity, map B, SD) (6P, 2018)	0.324	0.369	0.281	0.278	0.322	0.264	0.341
88	Mean Grade 8 residual of students with a Grade 7 grade below 30% (SD) (6P, 2018)	0.166	0.153	0.209	0.198	0.173	0.127	0.149
89	Mean Grade 8 residual of students with a Grade 7 grade of 80% or more (SD) (6P, 2018)	0.177	0.156	0.208	0.176	0.166	0.192	0.214
90	Lowest correlation of band-slope levels with baseline levels (6P, min over 12 fits)	0.997	0.999	0.998	0.998	0.998	0.997	0.998
91	SD of the calibrated learning level (6P, 2018, mean of 6 fits)	0.376	0.418	0.405	0.333	0.346	0.354	0.437
92	SD of the calibrated learning level (6P, 2022)	0.368	0.446	0.384	0.328	0.338	0.348	0.412
93	Joint R^2 of school characteristics for the persistent wedge (%) (GT matched schools)	—	37.1	19.7	9.7	11.4	—	—
94	Size-weighted correlation of the wedge with the raw Grade 7 school mean (6P, weighted mean over 12 fits)	-0.58	-0.463	-0.52	-0.655	-0.654	-0.642	-0.511

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
95	Lowest SD of the wedge over the 12 province-cohort fits (6P, range over 12 fits)	0.43	0.471 / 0.468	0.446	0.426	0.449	0.439	0.475
96	Highest SD of the wedge over the 12 province-cohort fits (6P, range over 12 fits)	0.49	0.471 / 0.468	0.467	0.44	0.45	0.465	0.486
97	Kane–Staiger signal share of wedge variance (6P, 2018)	0.84	0.946	0.792	0.844	0.826	0.817	0.861
98	Kane–Staiger signal share of wedge variance (6P, 2022)	0.866	0.946	0.817	0.859	0.872	0.874	0.88
99	Split-half correlation of the wedge (6P, 2018, mean of 6 fits)	0.69	0.884	0.606	0.702	0.693	0.61	0.72
100	Spearman–Brown projection of the split-half correlation (6P, 2018, mean of 6 fits)	0.817	0.939	0.755	0.825	0.818	0.758	0.837
101	Lowest correlation of the parametric alternatives with the baseline wedge (6P, both cohorts)	0.964	0.987	0.945	0.973	0.967	0.962	0.968
102	Lowest correlation of the baseline wedge with demographic-control wedges (2022 value printed) (GT, 2022)	0.998	0.9993	0.997	0.998	0.999	0.998	0.998
103	Correlation of the baseline wedge with the quintile-demeaned wedge (2018 and 2022) (6P, 2018 and 2022)	0.96 / 0.934	0.9171 / 0.8213	0.959 / 0.932	0.976 / 0.963	0.98 / 0.967	0.974 / 0.953	0.918 / 0.912
104	Share of noise-free wedge variance recurring in the other cohort (%) (6P, both cohorts)	52	66	53	52	41	51	50
105	Share of noise-free wedge variance specific to the cohort (%) (6P, both cohorts)	45	33	43	46	56	46	47
106	Share of noise-free wedge variance specific to the Grade 8 term (%) (6P, both cohorts)	3	2	3	3	4	4	2
107	Precision-weighted 2018–2022 correlation of calibrated levels (6P, schools fitted in both cohorts)	0.712	0.889	0.679	0.624	0.583	0.569	0.726
108	Spearman 2018–2022 correlation of calibrated levels (6P, schools fitted in both cohorts)	0.465	0.758	0.398	0.392	0.379	0.42	0.533
109	2018–2022 correlation of raw Grade 7 school means (6P, schools fitted in both cohorts)	0.386	0.441	0.412	0.446	0.271	0.391	0.321
110	Correlation of the 2018 raw mean with the 2022 calibrated level (GT, schools fitted in both cohorts)	0.241	0.316	0.313	0.152	0.129	0.209	0.249
111	Share of schools changing quintile of level between cohorts (%) (6P, schools fitted in both cohorts)	65.9	50.5	68.6	67.6	70	67.5	64.5
112	Schools with both an English and a mathematics wedge (6P)	8,732	842	2,492	1,951	1,959	767	721
113	Disattenuated correlation of English and mathematics wedges (6P)	0.601	0.667	0.591	0.62	0.556	0.581	0.677
114	Within-district share of wedge variance, precision-weighted (%) (6P, both cohorts)	94.41	89.2	96.4	94.8	98.4	94.3	98.2
115	Pearson correlation of NSC Mathematics with the raw Grade 7 grade (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.313	0.411	0.27	0.308	0.332	0.222	0.338
116	Pearson correlation of NSC Mathematics with the calibrated Grade 7 grade (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.389	0.54	0.345	0.355	0.369	0.34	0.373

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
117	Pearson correlation of NSC Mathematical Literacy with the raw Grade 7 grade (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.248	0.299	0.243	0.228	0.234	0.19	0.326
118	Pearson correlation of NSC Mathematical Literacy with the calibrated Grade 7 grade (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.398	0.574	0.377	0.333	0.318	0.378	0.469
119	NSC Mathematics gap, hardest vs easiest third of primaries, raw Grade 7 grade in the 60s (points) (6P, 2020 cohort)	5.7	8.7	6	5.6	5.8	4.1	2.3
120	Within-R^2 of NSC Mathematics, secondary-school FE, on the raw Grade 7 grade (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.152	0.188	0.147	0.147	0.175	0.128	0.157
121	Within-R^2 of NSC Mathematics, secondary-school FE, on the calibrated score (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.191	0.215	0.2	0.184	0.203	0.179	0.169
122	Within-R^2 of NSC Mathematics, secondary-school FE, gain from freeing the weights (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.005	0.006	0.002	0.004	0.007	<0.001	0.009
123	Within-R^2 of NSC Mathematical Literacy, secondary-school FE, on the raw Grade 7 grade (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.160	0.198	0.15	0.14	0.161	0.167	0.218
124	Within-R^2 of NSC Mathematical Literacy, secondary-school FE, on the calibrated score (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.204	0.248	0.2	0.18	0.198	0.207	0.257
125	Within-R^2 of NSC Mathematical Literacy, secondary-school FE, gain from freeing the weights (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.003	0.002	0.002	0.002	0.002	0.004	0.005
126	Wedge coefficient on sitting the NSC (participation LPM) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.340	0.302	0.332	0.362	0.362	0.333	0.321
127	Own Grade 7 grade coefficient on sitting the NSC (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.196	0.19	0.189	0.19	0.206	0.21	0.192
128	Share of the cohort sitting the NSC (base rate) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.396	0.454	0.415	0.404	0.366	0.395	0.337
129	Wedge coefficient on the NSC Mathematics score (M1) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.697	0.689	0.73	0.739	0.645	0.739	0.473
130	Wedge coefficient on the NSC Mathematical Literacy score (M1) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.751	0.81	0.749	0.784	0.692	0.679	0.746
131	Own Grade 7 grade coefficient on the NSC Mathematics score (M1) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.458	0.534	0.446	0.45	0.463	0.457	0.408
132	Own Grade 7 grade coefficient on the NSC Mathematical Literacy score (M1) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.463	0.522	0.458	0.436	0.417	0.477	0.508
133	1/beta1, the ratio the calibration implies (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1.91	1.815	1.901	2.113	1.988	1.758	1.896

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
134	Wedge-to-own-grade ratio on sitting the NSC (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1.74	1.595	1.76	1.902	1.762	1.588	1.675
135	Wedge-to-own-grade ratio (lambda/delta), NSC Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1.52	1.291	1.639	1.643	1.395	1.618	1.159
136	Wedge-to-own-grade ratio (lambda/delta), NSC Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1.62	1.551	1.637	1.796	1.658	1.425	1.47
137	Participation ratio, 2020 wedge instrumented with the 2018 and 2022 wedges (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1.95	1.627	1.955	2.329	1.962	1.812	1.753
138	Share of the 2020 calibration panel linked to NSC records (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	63.7	38.024	71.993	63.944	77.947	70.085	68.009
139	Link rate for students from Quintile 1 primaries (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	75.6	48.392	85.497	73.082	81.213	78.498	72.237
140	Link rate for students from Quintile 5 primaries (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	23.9	20.223	17.643	14.278	49.483	48.024	77.647
141	Quintile 5 minus Quintile 1 gap in raw Grade 7 school means (SD), no controls (6P, 2020)	0.269	0.188	0.558	0.328	0.244	-0.025	—
142	Quintile 5 minus Quintile 1 gap in raw Grade 7 school means (SD), no controls (6P, 2021)	0.255	0.145	0.603	0.317	0.37	0.017	—
143	Quintile 5 minus Quintile 1 gap in the calibrated level (SD), no controls (6P, 2020)	0.657	0.621	0.978	0.621	0.719	0.326	—
144	Quintile 5 minus Quintile 1 gap in the calibrated level (SD), no controls (6P, 2021)	0.706	0.679	1.047	0.67	0.813	0.398	—
145	Raw Quintile 5 minus Quintile 1 gap on the value-added every-term link (SD) (6P, 2020)	0.264	0.185	0.558	0.303	0.24	-0.023	—
146	Calibrated Quintile 5 minus Quintile 1 gap on the value-added every-term link (SD) (6P, 2020)	0.673	0.648	0.977	0.634	0.721	0.339	—
147	School-cohorts in the re-ranking (6P, 2018+2022 school-cohorts)	22,868	2,903	6,208	5,597	4,379	2,011	1,770
148	Share of the bottom raw quintile staying in the bottom level quintile (%) (6P, 2018+2022 school-cohorts)	22.4	27	22.1	22.6	20.1	20.3	23.4
149	Share of the top raw quintile staying in the top level quintile (%) (6P, 2018+2022 school-cohorts)	29.4	41.6	27.8	27	24.3	30.7	33.1
150	Share of the middle raw quintile staying in the middle (%) (6P, 2018+2022 school-cohorts)	20.4	21.5	20.3	19.8	19.6	20.9	22
151	Share of school-cohorts moving two quintiles or more (%) (6P, 2018+2022 school-cohorts)	44.7	38.1	44.7	46.6	46.6	45.5	43.3
152	Median rank shift from raw mean to calibrated level (places) (GT, 2018+2022 school-cohorts)	542	321	833.75	783	606	268	227.5
153	Share of school-cohorts moving two deciles or more (%) (6P, 2018)	68.1	60.7	69.7	69.7	70.6	65.9	66.3

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
154	Share of school-cohorts moving two deciles or more (%) (6P, 2022)	69.1	63.7	68.5	70.4	70.8	71.2	69.4
155	R^2 of the absolute rank shift on log school size (GT, 2018+2022 school-cohorts)	0.036	0.002	0.009	0.004	0.003	<0.001	0.002
156	Rank correlation of the absolute rank shift with school size (GT, 2018+2022 school-cohorts)	-0.176	-0.026	-0.091	-0.056	-0.049	-0.01	-0.042
157	School-cohorts with every characteristic recorded (production function) (6P, 2018+2022 pooled)	11,977	2,332	1,738	2,907	2,500	1,244	1,256
158	Quintile 5 coefficient on raw grades in Grade 8 units, no over-age control (SD) (6P, 2018+2022 pooled)	0.242	0.142	0.572	0.279	0.348	0.124	—
159	Quintile 5 coefficient on the calibrated level, no over-age control (SD) (6P, 2018+2022 pooled)	0.681	0.487	1.053	0.613	0.835	0.421	—
160	Raw share of the calibrated Quintile 5 coefficient, no over-age control (6P, 2018+2022 pooled)	0.356	0.292	0.543	0.455	0.417	0.294	—
161	Quintile 5 coefficient on raw grades in Grade 8 units, no over-age control (SD) (6P, 2018+2022)	0.242	0.142	0.572	0.279	0.348	0.124	—
162	Quintile 5 coefficient on the calibrated level, no over-age control (SD) (6P, 2018+2022)	0.681	0.487	1.053	0.613	0.835	0.421	—
163	Raw share of the calibrated Quintile 5 coefficient, no over-age control (6P, 2018+2022)	0.356	0.292	0.543	0.455	0.417	0.294	—
164	Log class size coefficient on raw grades in Grade 8 units, no over-age control (6P, 2018+2022 pooled)	-0.066	-0.205	0.002	-0.081	-0.031	-0.031	-0.083
165	Log class size coefficient on the calibrated level, no over-age control (6P, 2018+2022 pooled)	-0.127	-0.504	-0.067	-0.098	-0.041	-0.221	-0.06
166	Over-age coefficient on raw grades in Grade 8 units (SD) (6P, 2018+2022 pooled)	-0.244	-0.281	-0.149	-0.216	-0.265	-0.32	-0.261
167	Over-age coefficient on the calibrated level (SD) (6P, 2018+2022 pooled)	-0.089	-0.083	-0.098	0.01	-0.072	-0.092	-0.132
168	Raw share of the calibrated Quintile 5 coefficient, with over-age control (6P, 2018+2022 pooled)	0.266	0.222	0.488	0.271	0.345	0.248	—
169	Raw share of the calibrated Quintile 5 coefficient, with over-age control (6P, 2018+2022)	0.266	0.222	0.488	0.271	0.345	0.248	—
170	Romano–Wolf adjusted p of the no-fee indicator (GT, 2018+2022 pooled)	—	0	0	0.008	0	—	—
171	Romano–Wolf adjusted p of the school's own poverty quintile (GT, 2018+2022 pooled)	—	0.008	0.01	0	0.25	—	—
172	Standardised no-fee coefficient on the wedge, bivariate (6P, 2018+2022 pooled)	0.32	0.49	0.3	0.174	0.281	0.319	0.376
173	Standardised no-fee coefficient with own quintile, resources and mechanical controls (GT, 2018+2022 pooled)	—	0.426	0.314	0.077	0.259	—	—

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
174	No-fee gradient retained under interacted slopes, low end (ratio) (GT, 2018+2022 pooled)	—	0.99	0.95	1	0.93	—	—
175	No-fee gradient retained under interacted slopes, high end (ratio) (GT, 2018+2022 pooled)	—	1.01	1	1	1	—	—
176	Schools matched across cohorts (Shapley sample) (GT matched schools)	9,918	1,280	2,720	2,234	2,026	864	794
177	Shapley share of the joint R^2: downstream market structure (%) (GT matched schools)	—	38.3	27.2	17.7	32.3	—	—
178	Shapley share of the joint R^2: resources and structure (%) (GT matched schools)	—	28.8	40.8	42.2	14	—	—
179	Shapley share of the joint R^2: accountability exposure (%) (GT matched schools)	—	28.4	29.1	15.6	31.2	—	—
180	Student-subject-cohort observations in the revealed-preference test (GT movers, 2018+2022)	1,042,656	221,093	166,183	259,632	193,734	108,666	93,348
181	Origin schools in the test (GT movers, 2018+2022)	12,050	1,371	3,257	3,262	2,198	1,044	918
182	Destination schools in the test (GT movers, 2018+2022)	6,535	883	1,379	1,797	1,404	612	460
183	Raw Grade 7 mean on destination peers' Grade 7 mean (GT movers, 2018+2022)	0.411	0.310	0.296	0.397	0.512	0.457	0.423
184	Raw Grade 7 mean on peers' Grade 7 mean, own feeder left out (GT movers, 2018+2022)	0.108	0.109	0.042	0.08	0.116	0.044	0.103
185	Wedge on peers' Grade 8 mean, own feeder left out (6P movers, 2018+2022)	0.064	0.136	0.012	0.012	0.045	0.089	0.056
186	Movers in the published-matric test (GT movers, 2018+2022)	972,272	209,666	140,706	247,321	188,284	100,628	85,692
187	Destination secondaries in that test (GT movers, 2018+2022)	5,433	756	859	1,609	1,313	496	401
188	Wedge on the destination's published NSC Mathematics mean (6P movers, 2018+2022)	0.104	0.144	0.095	0.093	0.087	0.101	0.042
189	t-statistic of that coefficient (6P movers, 2018+2022)	13.227	7.1	5.615	5.975	4.68	4.589	1.413
190	Raw Grade 7 mean on the destination's published NSC Mathematics mean (6P movers, 2018+2022)	-0.009	0.049	-0.032	-0.043	-0.045	-0.013	0.03
191	Wedge coefficient over raw Grade 7 coefficient (GT movers, 2018+2022)	-11.762	3	-3	-2.145	-1.944	-7.906	1.408
192	Lower 95% bound of the wedge coefficient (6P movers, 2018+2022)	0.088	0.104	0.062	0.062	0.051	0.058	-0.016
193	Upper 95% bound of the wedge coefficient (6P movers, 2018+2022)	0.119	0.184	0.129	0.123	0.124	0.144	0.101
194	Lower 95% bound of the raw coefficient (6P movers, 2018+2022)	-0.022	0.019	-0.059	-0.071	-0.075	-0.052	-0.017
195	Upper 95% bound of the raw coefficient (6P movers, 2018+2022)	0.005	0.078	-0.005	-0.016	-0.015	0.027	0.077
196	Cross-cohort wedge on peers' Grade 8 mean (6P movers, 2018+2022)	0.068	0.108	0.041	0.04	0.065	0.072	0.016

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
197	t-statistic of the cross-cohort Grade 8 coefficient (6P movers, 2018+2022)	9.341	5.5	2.716	2.916	4.183	3.186	0.71
198	Wedge on the destination-netted own-feeder-out outcome (6P movers, 2018+2022)	0.23	0.231	0.215	0.206	0.227	0.226	0.202
199	t-statistic of that coefficient (6P movers, 2018+2022)	28.789	11.3	14.691	13.292	13.443	10.535	8.95
200	Students in the value-added estimation (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	394,353	45,084	85,425	86,185	85,411	54,432	37,816
201	Secondaries with at least 20 sitters, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1,135	143	256	227	250	187	72
202	Secondaries with at least 20 sitters, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	1,486	239	317	353	203	221	153
203	Signal SD of the exit-school effects theta, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.344	0.332	0.356	0.393	0.252	0.385	0.313
204	Signal SD of the exit-school effects theta, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.321	0.221	0.303	0.353	0.294	0.397	0.323
205	Disattenuated rank agreement of raw- and calibrated-intake theta, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.877	0.919	0.885	0.905	0.761	0.878	0.946
206	Disattenuated rank agreement of raw- and calibrated-intake theta, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.867	0.846	0.848	0.93	0.784	0.881	0.808
207	Share of secondaries changing quintile of theta between raw and calibrated intakes, Mathematics (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.404	0.385	0.387	0.405	0.46	0.412	0.333
208	Share of secondaries changing quintile of theta between intakes, Mathematical Literacy (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.442	0.427	0.454	0.368	0.527	0.425	0.471
209	Raw Term 1 placebo dispersion over theta dispersion, raw intake, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.696	0.302	0.689	0.623	0.706	0.94	0.449
210	Raw Term 1 placebo dispersion over theta dispersion, raw intake, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.704	0.678	0.738	0.633	0.82	0.739	0.606
211	Calibrated Term 1 placebo dispersion over theta dispersion, calibrated intake, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.364	0.131	0.406	0.331	0.362	0.439	0.279
212	Calibrated placebo dispersion ratio, calibrated intake, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.434	0.423	0.474	0.413	0.483	0.44	0.314
213	Split-half reliability of theta, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.841	0.657	0.837	0.867	0.834	0.89	0.886

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
214	Lowest agreement with theta across intake-preserving specifications, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.972	—	0.946	0.989	0.929	0.976	0.982
215	Lowest agreement with theta across intake-preserving specifications, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.973	—	0.984	0.963	0.961	0.987	0.948
216	Agreement with theta under the median attrition scenario, Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.861	0.922	0.896	0.854	0.818	0.841	0.814
217	Agreement with theta under the median attrition scenario, Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.801	0.813	0.799	0.842	0.724	0.838	0.802
218	Placebo's projection share of theta's signal variance (%), Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.018	0.034	0.008	0.039	0.041	0.038	0.046
219	Placebo's projection share of theta's signal variance (%), Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.028	0.027	0.01	0.029	0.119	0.052	0.044
220	One-for-one worst case share (%), Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.186	0.082	0.215	0.133	0.206	0.262	0.136
221	One-for-one worst case share (%), Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.296	0.282	0.34	0.231	0.365	0.316	0.251
222	Share of secondaries in the same band after the sorting adjustment (%), Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.951	0.944	0.957	0.95	0.956	0.941	0.958
223	Share of secondaries in the same band after the sorting adjustment (%), Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.951	0.992	0.943	0.935	0.951	0.941	0.954
224	Share in the same band under the Oster bound (%), Mathematics (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.78	0.713	0.77	0.784	0.804	0.791	0.819
225	Share in the same band under the Oster bound (%), Mathematical Literacy (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.777	0.699	0.785	0.819	0.818	0.751	0.771
226	Share of secondaries changing band in at least 20% of bootstrap draws, Mathematics (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.402	0.329	0.406	0.383	0.476	0.396	0.347
227	Share of secondaries changing band in at least 20% of bootstrap draws, Mathematical Literacy (%) (2020 cohort (Gr 7 2020 → NSC 2025), 6P)	0.386	0.372	0.407	0.382	0.404	0.348	0.399
228	Rank correlation with the known effects under sorting at observed density (Monte Carlo, generated data)	—	—	—	—	—	—	—

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
229	Rank correlation in the sparsest mobility scenario (Monte Carlo, generated data)	—	—	—	—	—	—	—
230	Within-feeder correlation of the Grade 7 grade with the destination effect (empirical mixing proxy) (6P, 12 province-cohort strata pooled)	0.021	0.015	-0.002	0.013	0.054	0.029	0.027
231	Proxy value at which endogenous mobility would bias the variance by 5 pp (Monte Carlo, generated data)	—	—	—	—	—	—	—
232	Ratio of that threshold to the observed proxy (Monte Carlo, generated data)	—	—	—	—	—	—	—
233	Threshold heaping, bottom over top fifth of schools by raw mean (ratio) (GT, 2018+2022 pooled)	5.173	5.6	4.619	6.241	5.679	4.032	4.114
234	Lowest median Term 4/Term 1 heaping ratio, Grades up to 9 (DBE extract 2021–24, its provinces pooled)	1.187	0.907	1.141	1.075	1.168	1.184	0.814
235	Highest median Term 4/Term 1 heaping ratio, Grades up to 9 (DBE extract 2021–24, its provinces pooled)	1.354	1.278	1.562	1.372	2.497	1.343	3.03
236	Median Term 4/Term 1 heaping ratio, Grade 12 (DBE extract 2021–24, its provinces pooled)	0.988	0.978	1.064	0.904	1.008	0.965	1.118
237	Median Term 4/Term 1 heaping ratio, Grade 10 (DBE extract 2021–24, its provinces pooled)	1.041	1.121	1.227	0.972	1.027	0.99	1.202
238	Median Term 4/Term 1 heaping ratio, Grade 11 (DBE extract 2021–24, its provinces pooled)	1.065	1.093	1.274	0.945	1.029	1.014	1.063
239	Heaping index injected at 40 in the simulation (Monte Carlo, generated data)	—	—	—	—	—	—	—
240	Rank correlation with the truth under observed heaping (Monte Carlo, generated data)	—	—	—	—	—	—	—
241	Rank correlation with the truth with no heaping (Monte Carlo, generated data)	—	—	—	—	—	—	—
242	Shift in the slope between heaped and clean grades (Monte Carlo, generated data)	—	—	—	—	—	—	—
243	Largest rank-correlation deficit under quality-correlated heaping (Monte Carlo, generated data)	—	—	—	—	—	—	—
244	Largest slope change under quality-correlated heaping (%) (Monte Carlo, generated data)	—	—	—	—	—	—	—
245	Largest change in beta1 from de-heaping (% of beta1) (6P, max over 12 fits)	2.12	1.038	2.116	1.258	1.618	1.269	1.096
246	Lowest correlation of raw and de-heaped wedges (6P, min over 12 fits)	0.997	1	0.997	0.999	0.998	0.998	0.998
247	Number of comparisons in the single-fixed-effect check (DBE extract, comparisons by grade × subject × province)	—	—	—	—	—	—	—
248	Lowest correlation of the single FE with the sum of intake and exit effects (DBE extract, comparisons by grade × subject × province)	0.9944	—	0.996	0.994	0.995	0.994	0.996

(continued)

	What the main text cites	Six provinces	GT	EC	KZN	LP	MP	NW
249	Highest correlation of the single FE with that sum (DBE extract, comparisons by grade × subject × province)	0.9999	—	1	1	1	1	1
250	Lowest correlation of the single FE with the exit effect alone (DBE extract, comparisons by grade × subject × province)	0.015	—	0.13	0.294	0.116	0.197	0.02
251	Highest correlation of the single FE with the exit effect alone (DBE extract, comparisons by grade × subject × province)	0.607	—	0.52	0.576	0.462	0.607	0.386
252	Spearman–Brown reliability of the two-way exit-school effect, Mathematics, Grade 8 intake (DBE extract 2021–24, its provinces pooled)	0.377	—	0.352	0.487	0.368	0.493	0.049
253	Schools behind the exit-effect reliability (DBE extract 2021–24, its provinces pooled)	1,320	—	177	487	279	206	75
254	Spearman–Brown reliability of the wedge (same statistic as the split-half projection) (GT, 2018)	0.817	0.939	0.755	0.825	0.818	0.758	0.837
H1	Grades just above 40 over grades just below it (heaping step, Grade 7 mathematics) (6P, every province and year (range))	two to seven times	3.2–7	2.7–10.8	1.9–8.4	2.2–6.2	1.8–5.4	1.9–3.4
H2	Grade 12 mathematics count at 40 against its smooth counterfactual (DBE extract 2021–24, its provinces pooled)	within 5 per cent	3.823	3.976	6.449	6.405	3.838	22.365
H3	Grade 7 mathematics count at 40 against its smooth counterfactual (DBE extract 2021–24, its provinces pooled)	about two to three times	2.5–2.6	1.7–2.2	1.9–2.8	2.2–3.0	1.7–2.4	1.5–2.1

Notes: Each row is a number the main text cites, in the order it first appears, with the sample the main text gives it for in parentheses (6P for the six provinces, GT where the main text names a number as Gauteng's). The first column is the six provinces, each fitted on its own and then combined as §2.4 of the main text describes; the others give the same quantity for each province's own estimation sample, fitted on its own. A dash marks a number with no province equivalent (a count over provinces, a simulation result or a literature figure) or a cell that cannot be computed: too few schools enter it, the DBE extract holds no Gauteng records for it, its specification controls for a rural flag the school register does not record in Mpumalanga or North West, or, for Grade 7 enrolments, the province counts exist only in the raw records. Two numbers separated by a slash are the 2018 and 2022 cohorts; two joined by a dash are a range. Numbers are rounded to three decimals, and counts are given in full. Three limits of the data sit inside the six-province numbers. The school register's no-fee flag is mis-coded in three province-years, so for the Eastern Cape's 2018 cohort and the Mpumalanga and North West 2022 cohorts it is taken from the nearest year that is coded (2020, 2018 and 2018). The matric test's two counts (rows 186 and 187) are computed on a newer version of the panel than the main text's, which has 25 more movers and one more destination school. In rows 93 and 113 the Gauteng cell comes from an older version of the school estimates than the other cells, and differs in the last printed digit from the same computation on the version they use. Sample: each row's own, in parentheses after it.

Appendix F: Precision and persistence of the wedge

A school's wedge is estimated from one cohort of its students, so part of it could be estimation noise, and part could belong to that cohort rather than to the school. Estimation noise is small, though not nil, and about half of the wedge lasts. By the leave-out correction, estimation noise is 16.1% and 13.4% of the wedge's variance in 2018 and 2022 (5.4% and 5.5% in Gauteng, where the network is densest), and two estimates from random halves of a school's students correlate at 0.69 in 2018 and 0.74 in 2022, averaged over the six provinces (0.88 and 0.88 in Gauteng). Net of estimation noise, about 52% of the wedge's variance recurs in the other cohort (66% in Gauteng), four years later and across the COVID-19 pandemic, though nothing in the design links the two cohorts. About 45% is specific to the cohort though shared by all four Grade 8 terms, and about 3% is specific to the Grade 8 term. The sections below, Gauteng's unless they say otherwise, give

the decomposition behind the within-cohort numbers, the correlations between the two cohorts' wedges, a middle cohort that separates a shock to one cohort from steady drift, the split of the wedge into noise, a lasting part and a part specific to the cohort, and how far schools revert between cohorts.

F.1: Reliability of the calibrated measure: variance decomposition and split-sample reproducibility

The main text states the reliability ratios and split-sample correlations of the calibrated school effect across the six provinces, with Gauteng's beside; this table carries Gauteng's full decomposition. The variance-decomposition rows give the Kane–Staiger signal share of the cross-school variance in $\hat{\alpha}_j$ (raw Grade 8 grade 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 9: **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 grade points. Row 3, Var(signal) = Row 1 minus Row 2 (floored at zero). Row 4, reliability ratio λ = Var(signal)/ 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 F.2.* ***Sample:*** *Gauteng, 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	–

F.2: 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 grades allow a partial decomposition, given at the end of this subsection. They are jointly unresolvable in this design, and the paper's contribution does not depend on resolving them.

In Gauteng, the precision-weighted Pearson cross-year correlation between the 2018 and 2022 wedges is 0.624 (Table 10; 0.477 across the six provinces), 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 10: 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, 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 9 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 grades 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 grade, which wedges calibrated from the same school's Term 1 to Term 3 grades do not share; that part is about 22% of the wedge's variance. (In 2022 the Term 4 grade is built from the earlier terms' grades, 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.

F.3: 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 in Appendix F.2, the statistic is the correlation of the school estimates $\hat{\alpha}_j$ across cohorts. The four-year rung reproduces Table 10: 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 matches a school component that persists across all three cohorts plus a component specific to each cohort. Across the 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.

F.4: Variance anatomy of the school-level grade–learning gap

Table 11: Variance decomposition of the school-level grade–learning gap (full table).

Component	GT (unwtd)	GT (wtd)	6P (unwtd)	6P (wtd)
(i) Estimation noise (mean SE²)	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(SE²) [Kane–Staiger].

F.5: 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 12: Dynamics of the school-level grade–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, × 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 coefficient indicates mean-reversion; this is partly mechanical noise-reversion. Significance: *** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.10.*

Appendix G: Tests of the exclusion restriction

The regression attributes the gap between two students who left different primary schools with the same Grade 7 grade to how hard the two primaries' assessments were. The same gap would arise if one primary's students learned more after leaving. Where the data can test this, the wedge passes. Between Terms 2 and 4, students from primaries with harder assessments gain 0.026 SD on their classmates for each SD of difference

in the wedges, less than the 0.037 SD that difficulty alone predicts (§3.3 of the main text; Appendix G.2), and the wedges keep their order whichever Grade 8 term anchors them and inside every group of schools (Appendix G.1, Appendix G.3). The size of a wedge still partly reflects the school's intake (Appendix G.4, Appendix G.5). Two conditions cannot be tested: the holiday between Grade 7 and Grade 8 has no assessment, and learning that Grade 7 did not test would show from the first Grade 8 term, just as difficulty does. The wedge's precision, persistence and prediction of later results would hold either way, so they do not settle it.

G.1: The Grade 8 anchor term

Table 13 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 province-cohort samples 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 G.2 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 13 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 13 reports the standard deviation of the wedge at each of the anchors of Appendix G.2, 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 grades 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. Across the six provinces, each calibrated on its own, every other term's wedge correlates with the Term 2 wedge at 0.964 or more in both cohorts once noise is removed, the figure the main text gives. 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 grade is in effect a year grade that contains the Term 1 grade, 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 13: **The Grade 8 anchor term does not change the ordering of schools.**

	Headline pool (1)	Sensitivity pool (2)
Panel A. This extract, Term 1 against Term 4 at the Grade 8 anchor		
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
Panel B. Grade 8 anchor ladder, Gauteng core		
SD of the wedge, Grade 8 Term 2 anchor	0.452	

(continued)

	(1)	(2)
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 province-cohort sample where both anchors estimate. Each row of Panel A is one statistic computed on the pair of school effects a province-cohort sample 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 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-cohort samples of the Grade 7 2021 and 2023 cohorts, and the sensitivity pool 8,551 schools over every further estimable province-cohort sample. Panel B, the paper's own Gauteng Mathematics panel of 2,562 school-cohorts carrying all four anchor grades.

G.2: 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 0.32 cross-school standard deviations lower across the six provinces, and 0.49 in Gauteng. 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 grading 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 grade one within-province standard deviation higher predicts between 0.46 and 0.56 of a standard deviation more on the Grade 8 grade 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 G.3 finds the same ordering in raw grades. Table 14 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 combinations it retains between 0.99 and 1.01 of its single-slope value. Imposing a common slope does not create the gradient.

Table 14: **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
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 15 re-estimates the wedge on the Grade 8 Term 2, Term 3, and Term 4 grades. 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 G.1).

Table 15: **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 grades.* ***Sample:*** *Gauteng Mathematics, 2,562 school-cohorts with all four grades, 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 grade down by the same amount, whatever their Grade 7 score. Table 16 fits the Grade 8 z-score on the Grade 7 z-score, a no-fee indicator, and their interaction, with destination-by-cohort fixed effects and standard errors 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 grading, with noisier no-fee Grade 7 grades 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 grading. 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 grade, 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 16: **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, with destination-by-cohort fixed effects; 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
Grade 7-by-no-fee interaction	-0.149	0.018	< 0.001

G.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 each province's cohort; for Gauteng, Table 17 reports the analogous check by re-fitting the baseline specification of §3.2 of the main text inside 12 subgroups defined by three school-level partitions (school mean Grade 7 grade quintile, national poverty quintile, and urban/rural status), alongside a school-size tercile check.

Table 17: 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) (points)	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 grades inside one subgroup, so $\hat{\beta}_1$ is Grade 8 grade points per Grade 7 grade point, not the within-province z-score slope of §3.2 of the main text. Sample: Gauteng, 2018 and 2022 cohorts.

Across the 12 subgroups the within-stratum pass-through $\hat{\beta}_1$, in Grade 8 grade points per Grade 7 grade 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 subgroups 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-grade 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.

G.4: 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 grades 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 grade-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 18: 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.) (points)	SD(Āp) (points)
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, 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.

G.5: What the noisy Grade 7 grade does to the wedge

The noisy Grade 7 grade biases the wedge as well as the slope. Within a school, the regression discounts each student's grade for its noise, which is why $\hat{\beta}_1$ converges to $\lambda\rho$. A school's mean grade needs no such discount, because the noise averages out over its students, yet the wedge subtracts the discounted slope times that

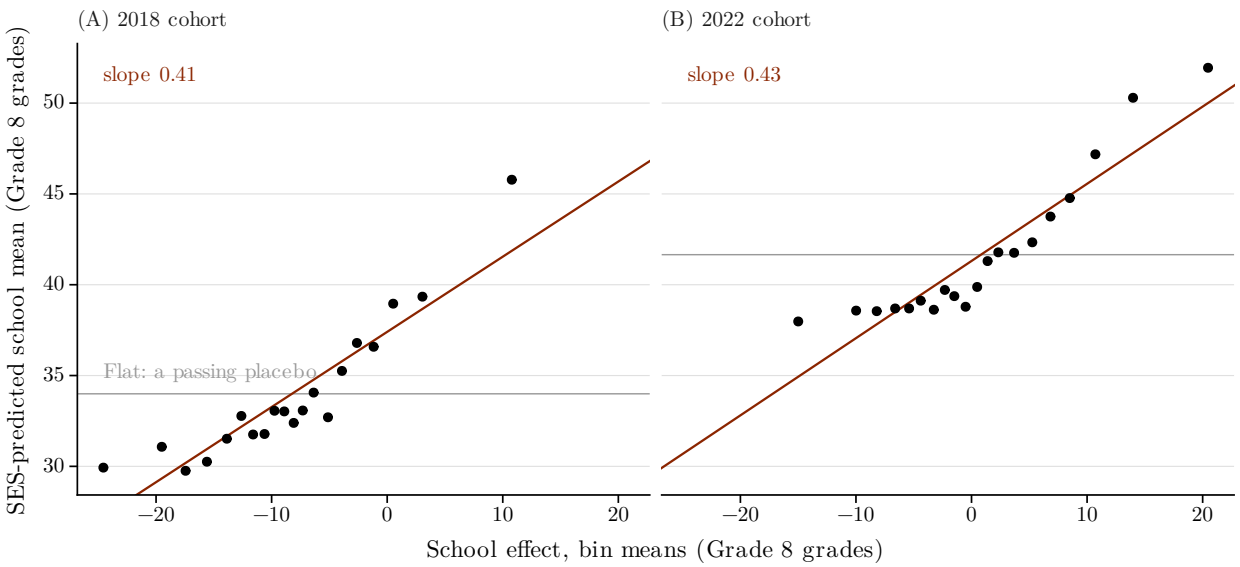

Figure 1: The SES-predicted mean $\bar{A}_j^p$ against $\hat{\alpha}_j$, as means of equal-count bins of schools, one panel per cohort, with the fitted line of each cohort and the flat line a passing placebo would trace. *Sample:* Gauteng, 2018 and 2022 cohorts.

mean. To first order and net of destination effects (the caveat of Appendix B.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 grade ($\rho = 1$) both distortions vanish.

How much this matters can be read off the data without fixing λ. At any slope b the wedge is exactly $\hat{\alpha}_j(b) = \mathrm{FE}_j(\text{Grade 8}) - b\,\mathrm{FE}_j(\text{Grade 7})$, with both feeder effects net of destination effects, and under the model $b = \lambda$ removes the intake term. Figure 2 recomputes, for Gauteng, how much easier no-fee primaries' assessments 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 grades 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 grades 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 Gauteng's instrumental-variables slope, 0.760 (2018) and 0.734 (2022), its row of Table 3 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. Across the six provinces, the figures the main text gives, the contrast rises from 0.41 at OLS to 0.47 at the instrumental-variables slope, and the intake term supplies 13–14% of it.

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

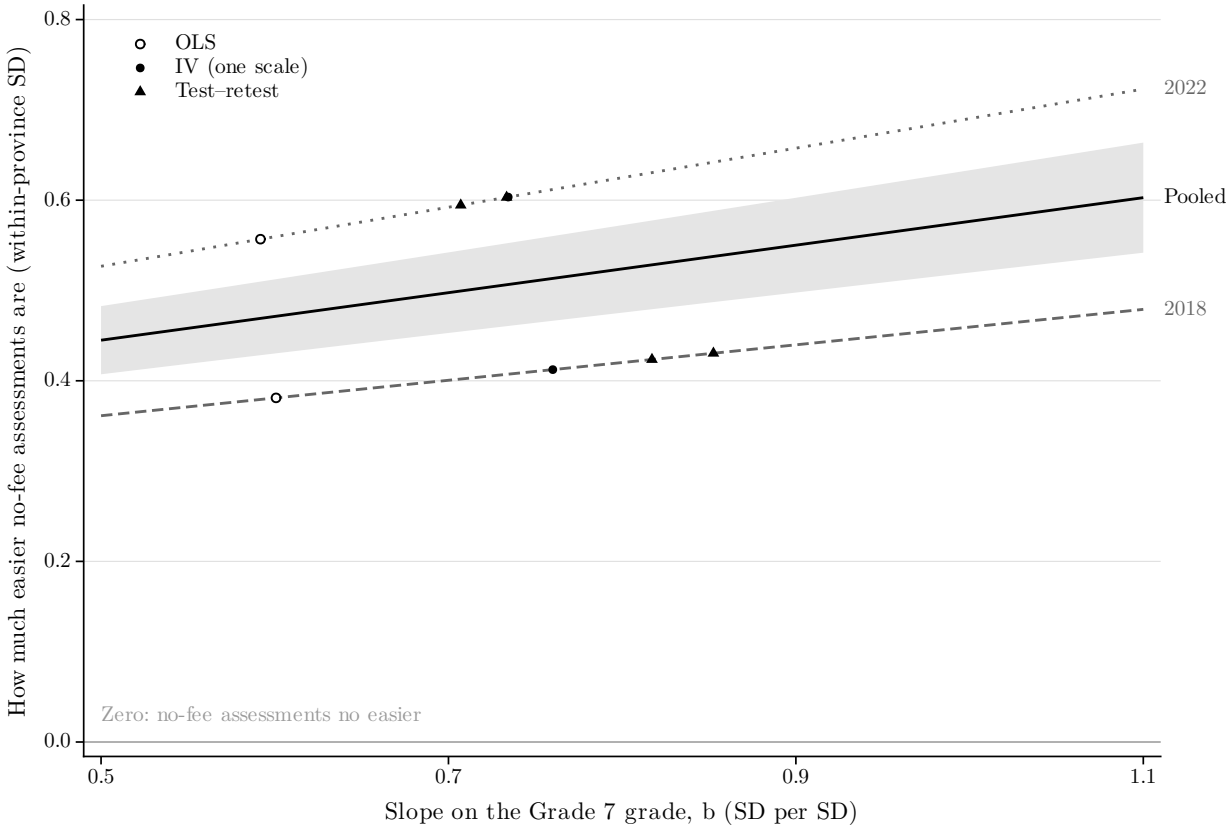

Figure 2: **Correcting for the noisy Grade 7 grade widens the no-fee gap.**

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 grade as $\hat{\alpha}_j(b) = \mathrm{FE}_j(\text{Grade 8}) - b\,\mathrm{FE}_j(\text{Grade 7})$, 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 grade; slopes are in SD of the Grade 8 grade per SD of the Grade 7 grade. 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, 2018 and 2022 cohorts; the contrast covers 2,871 school-cohorts with a fee status.

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 B, and four features of the data qualify it.

First, the instrument's sample. The panel holds no English Home Language grade, so the IV is fitted only on students with an English FAL grade: 60% of Gauteng'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 Gauteng's students (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 grades 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' grades are noisier, the shrinkage differs by school and the split above does not capture it; this is not tested here. Appendix G.2 finds the Grade 7 slope flatter among no-fee primaries, which noisier no-fee grades would produce, among other channels.

Appendix H: Tests of exogenous mobility

The design compares leavers of one primary school who share a Grade 7 grade but enter different secondary schools. If those bound for one secondary school were already improving faster, or knew more than their grades showed, the secondary schools' effects would absorb the difference, and the wedges with them. Students do sort, but mainly on the Grade 7 grade the regression holds fixed. Leavers of one primary bound for the top and for the bottom quarter of secondary schools gained within 0.012 SD of each other over the four Grade 7 terms, on either ranking of the secondary schools, where the wedges spread by 0.453 SD in 2018 (Appendix H.1). Sorting on gender and over-age is weaker, though beyond chance (Appendix H.2); a student fixed effect reorders the primary schools appreciably only in 2018 (Appendix H.7); and the data show less than half the sorting on students' own grades that would bias the wedges' variance by 5% (Appendix H.8).

H.1: Grades before the move

Figure 3 plots the grades 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).

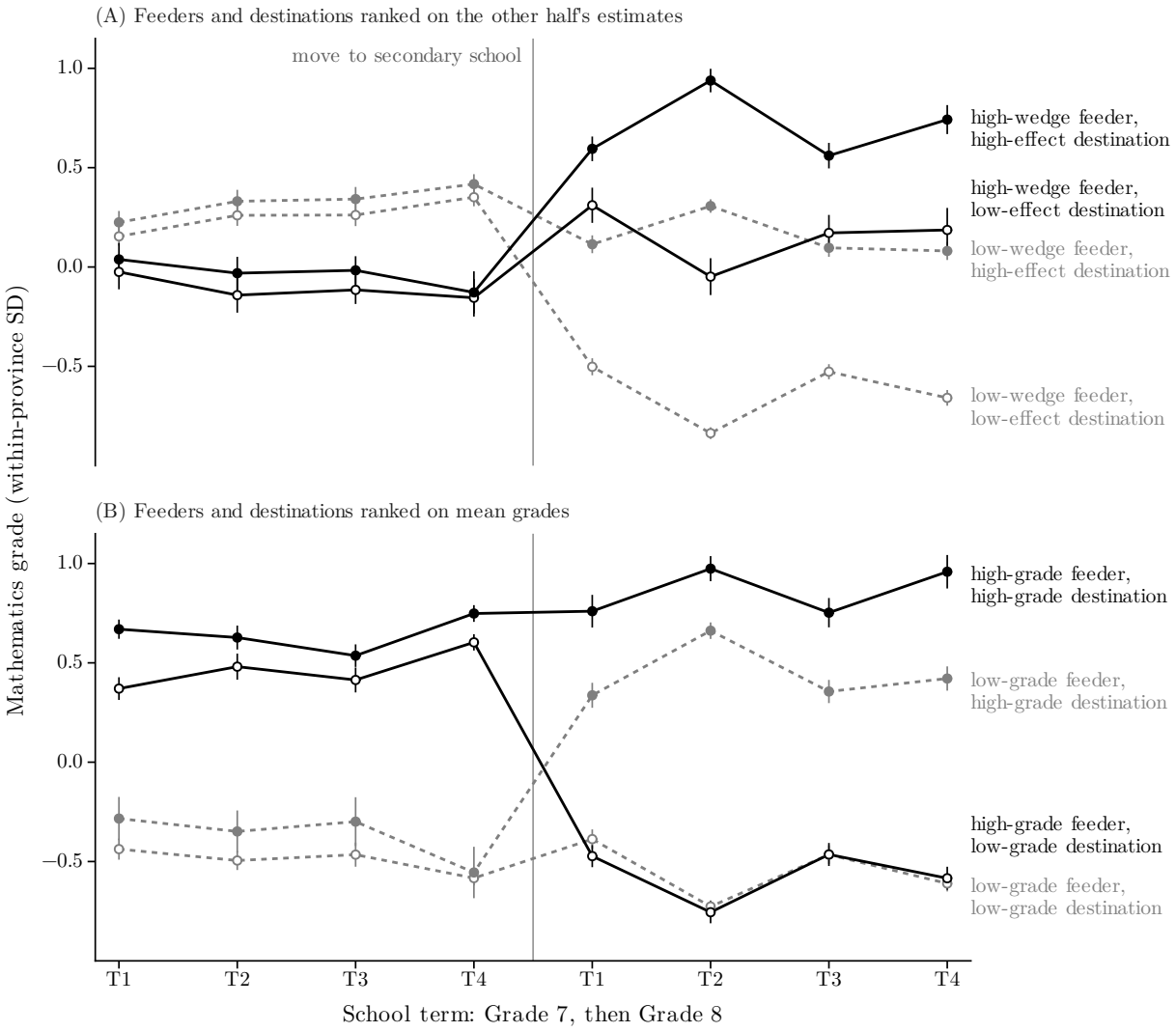

Figure 3: **Paths to high- and low-effect destinations ran nearly parallel before the move, by feeder quarter**

Notes: ***Each line follows the mean Mathematics grade, 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 grades 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 grade and destinations on their mean raw Grade 8 grade, each leaving the student's own grade out, the grouping the labour literature uses. Grades 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 grade, 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 J.1 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 grades, 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 grades, as in the standard labour diagnostic (Kline 2024, 141), students bound for high-grade destinations gain a few hundredths of an SD more before the move than their schoolmates bound for low-grade 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 J.1). The groups above are not held at one Term 4 grade, the grade 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 grade and asks whether those whose Term 4 grade 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 grade itself. Added to the model as a control, after removing its correlation with the Term 4 grade 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 grade is built from the three earlier terms, so the departure barely predicts the Grade 8 grade and the wedges do not move; the test has power only in 2018.

H.2: The sorting permutation test

This subsection and the four after it describe 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.

H.3: The permutation test: specification

For each feeder cohort (feeder school j, cohort year t) 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 p-value p_{jt} for each feeder cohort. Stacking p-values across all eligible feeder cohorts 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 grade**: the headline continuous measure of student ability. Grade 7 grades 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 grades 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 grades; "grades" 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 feeder cohort.

H.4: The permutation test: placebo construction

A common concern with F-tests by feeder cohort on bipartite mover networks is that in feeder cohorts 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 feeder cohorts. 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 F-test behaviour in sparse feeder cohorts by permuting student destinations within each feeder cohort 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 cohort. 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 and feeder cohort on each replicate. The resulting placebo distribution provides the reference against which the observed distribution is compared.

H.5: The permutation test: summary statistics

For each predetermined characteristic, we report:

- The share of feeder cohorts 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.

H.6: Sorting permutation test, observed vs placebo

Table 19: Within-feeder permutation test for non-random sorting into Grade 8 destinations.

Characteristic	Feeder cohorts	Share of feeder cohorts with p < 0.05				KS distance	
		Observed	Placebo	95% interval	Excess	Observed	Placebo
Grade 7 Term 4 grade	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 grade 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 feeder cohorts tested: 22,142. Placebo distribution: 200 within-feeder permutations preserving feeder-to-destination flow counts.

H.7: A student fixed effect

The student-fixed-effect version fits each student's eight term grades with the student's own effect, the effect of the school that set each assessment 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 20: **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 grades	817,988	1,026,096

Notes: The student-fixed-effect version models each of a student's eight term grades (Grade 7 and Grade 8, Terms 1 to 4) as the student's own effect, plus the effect of the school that set the assessment (the feeder in Grade 7, the destination cohort 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 grade; 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 grades and the baseline wedges in SD of Grade 8 grades, 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: outputs/tables/mathematics/tab_akm_appendix_student_fe.csv, built from analyses/seminar_checks/outputs/tables/tabQ14_compare.csv and tabQ14_sample.csv; each number's origin is listed 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 grades; 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 20). The difference lies in the Grade 7 grades, not in the student effect as such or in the Grade 8 grades. The student effect averages a student's four Grade 7 term grades, 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 grade carries a part specific to the feeder and the term. Across feeders, the mean grade 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 grades run high against their earlier terms do not send students who do better in Grade 8 against those earlier grades (slope 0.04, standard error 0.06, where learning would carry over at about 0.75). In 2022 the Term 4 grade is in effect a year grade 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 grade, the grade that decides promotion; in 2018 that difficulty includes a part specific to the term.

H.8: Endogenous-mobility bounding against the Andrews/KSS 5% threshold

The endogenous-mobility grid in Table 21 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 grade (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 feeder cohorts in 12 subject-province-cohort strata, the empirical signature is 0.021 (stratum means from -0.01 to 0.06). At Gauteng's network density (Figure 4), 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 $\psi^{\star} \approx 0.119$, 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 4), not endogenous-mobility bias, and shrinks towards zero as school size grows. On Gauteng's 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 I.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 Table 21 documents for $\psi \geq 1$ is a stress test, not a description of the empirical setting.

Table 21: **Rankings survive every simulated threat; sorting and endogenous mobility distort the size of the estimates.**

Threat	Scenario	Rank correlation, base	Rank correlation, severe	Mean bias (points)	Quintile match	Verdict
Heaping	Heaping index 1 to 3	0.968	0.968	<0.001	74.7%	Negligible
Thin mobility	2 to 11 destinations, 15 to 80 students	0.968	0.876	<0.001	74.6%	Negligible at observed density
Endogenous mobility	Mobility strength 0 to 2, at Gauteng's density	0.963	0.897	RMSE 2.51 at strength 2	58.5%	Ranks stable; levels degrade
Non-random attrition	Random to extreme selection	0.960	0.960	<0.001	72.0%	Negligible
Sorting	Sorting strength 0 to 1, peer effects 0 to 0.15	0.953	0.985	<0.001 (RMSE 5.84 at strength 1)	79.4%	Ranks stable; levels degrade

Notes: Each row is one simulated threat, on a synthetic system of 200 primary and 50 secondary schools in which each primary sends 50 leavers to 8 secondary schools (11 in the two mobility rows; in Gauteng's data the median primary sends leavers to 12); the heaping row is the simulation of §5.3 of the main text. The two rank-correlation columns give the rank correlation of the estimated with the known effects with no threat (for thin mobility, at 11 destinations) and at the most severe setting. Mean bias is zero by construction; two rows add the root mean squared error. Quintile match is the share of schools placed in their known effect's quintile, at the base setting except for endogenous mobility and sorting. Endogenous mobility is fitted with the paper's regression; heaping, thin mobility and attrition with primary and class effects; sorting with primary effects only (§5.1 of the main text reports it with the paper's regression). Attrition cannot bias the estimates by construction. 445 to 500 replications per scenario. Sample: simulated data.

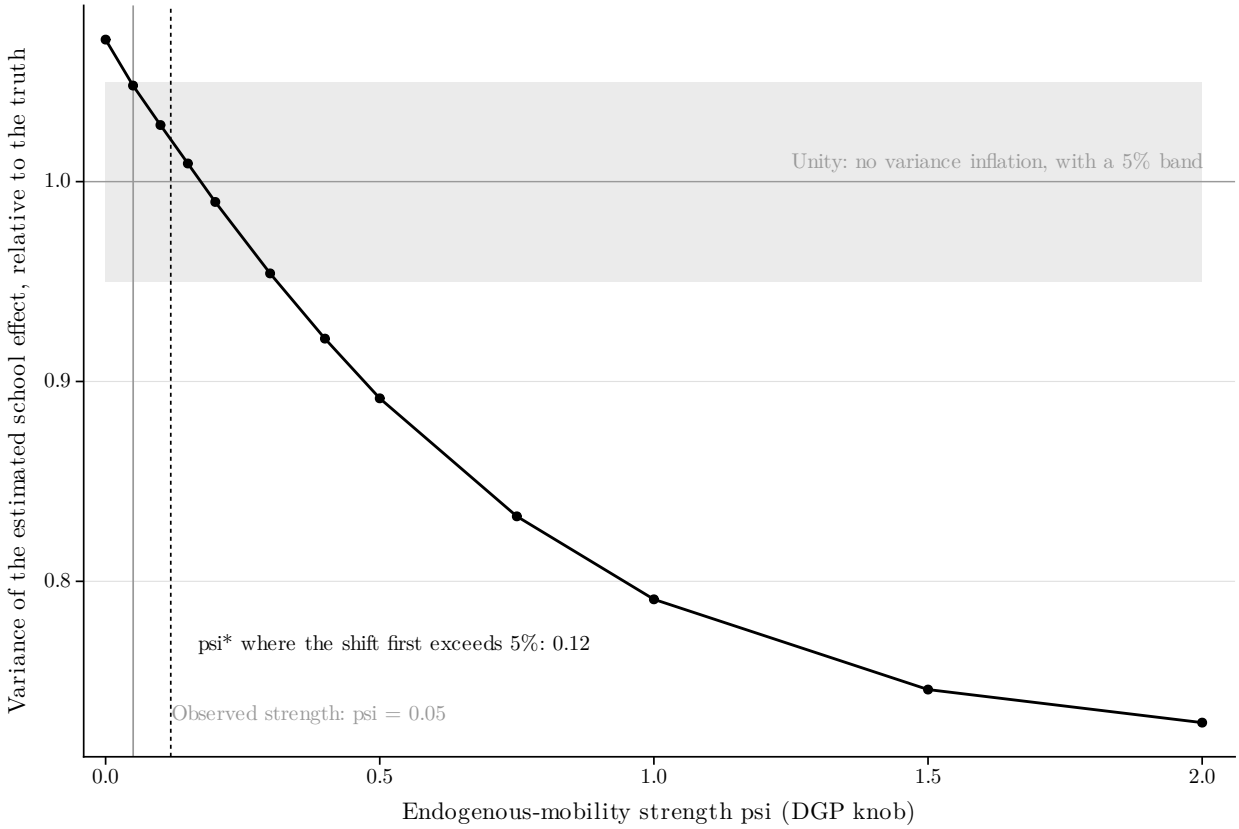

Figure 4: Bounding endogenous-mobility strength against the Kline et al. (2020) 5% threshold. Black curve: mean $\mathrm{Var}(\hat{\alpha})/\mathrm{Var}(\alpha)$ across 500 MC replications per ψ at the network density of Gauteng (11 destinations, 50 students/feeder). Grey band: $\pm 5\%$ deviation band. Solid grey line: ψ implied by the empirically-observed within-feeder $\mathrm{corr}(\mathrm{baseline}_{\mathrm{within}}, \hat{\gamma})$ of 0.021. Dotted black line: $\psi^{\star}$ 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:* simulated data; the empirical correlation is across the six provinces, 2018 and 2022 cohorts.

Appendix I: Attrition and limited mobility

The estimates use students who can be followed from Grade 7 into a Grade 8 class, in sets of schools that students' moves link. If the students who cannot be followed were concentrated in schools with easier or harder assessments, or if too few moves linked a school to the rest, the wedges would carry the selection or the noise. Neither does much. Across 22,868 school-years, measured within each province and cohort, a school's wedge and its linkage rate correlate at −0.034, and weighting by the inverse of the linkage rate shifts the mean wedge by 0.004 SD (Appendix I.4, Appendix I.5). Almost every primary school lies in the largest connected set of its province and cohort, and the few outside it are small (Appendix I.1, Appendix I.2); the bias from thin links is small, though not nil (Appendix I.3). Each province is its own connected set, so the wedges compare primary schools within a province, not across provinces.

I.1: The connected set

Wedges are comparable only within a set of schools that students' moves link, so each province and cohort is estimated on its largest connected set (Kline 2024, 126–27). In every row of Table 22, that set holds at least 98.8 per cent of the primary schools and 99.3 per cent of the linked students.

Table 22: **Almost every primary school sits in the largest connected set of its province and cohort**

Province	All primary schools	Largest connected set			Components
		Primary schools	Primary schools (%)	Students (%)	
Panel A. 2018 cohort					
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
Panel B. 2022 cohort					
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 sample, 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 sample 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 sample's primary schools and of its linked students; components count the separate sets in the sample'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: outputs/tables/mathematics/tab_akm_appendix_coverage.csv, built from outputs/tables/mathematics/connected_set_coverage.csv; each number's origin is listed in outputs/tables/mathematics/tab_akm_appendix_sources.csv. Sample: the National (6 provinces) sample, Mathematics, the 2018 and 2022 cohorts: every province and cohort, with 533,419 and 665,514 linked students and 11,047 and 11,924 primary schools.*

The leave-out correction of Appendix I.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 24 against Table 22).

I.2: Connected-set drop-out

Schools and students removed by the within-cohort connected-set restriction in Gauteng, 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 23: Connected-set drop-out: schools and students removed by the within-cohort connected-set restriction, summarised by observable partition (pooled 2018 + 2022).

		Schools dropped		Students dropped		
Partition	Test	n	(%)	n	(%)	p-value
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 Mathematics, pooled 2018 and 2022 cohorts.

I.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 24 reports, for Gauteng, 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 24: **Estimation noise is a small share of the variance of the wedge, but not nil**

		SD of the wedge			
Correction and weighting	Primary schools	Plug-in	Corrected	Noise share (%)	Precision
Panel A. 2018 cohort					
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

(continued)

Correction and weighting	Primary schools	Plug-in	Corrected	Noise share (%)	Precision
Errors shared within a link; schools weighted by students	1,409	0.446	0.432	6.0 (0.7)	Fail
Panel B. 2022 cohort					
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 grade. 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: outputs/tables/mathematics/tab_akm_appendix_noise.csv, built from analyses/wedge_noise_correction/outputs/tables/corrected_variance.csv and validation.csv; each number's origin is listed 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 (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 grade.

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 5). 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.

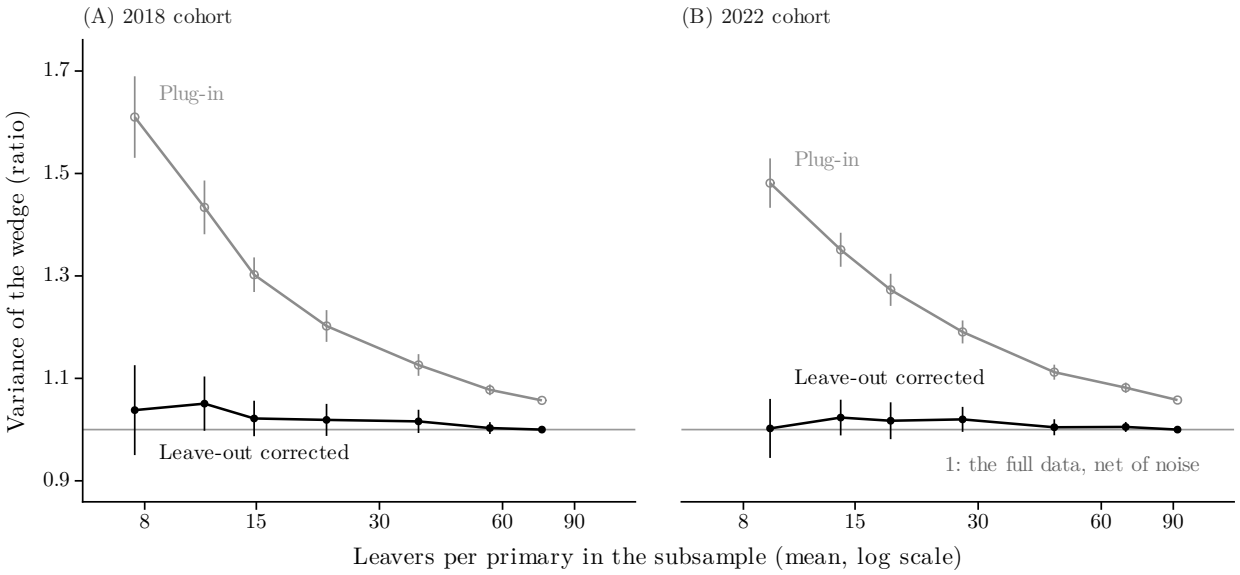

Figure 5: **Thin primaries inflate the plug-in variance, not the corrected one**

***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 25 gives the decomposition the two-way fixed-effects literature reports first (Abowd et al. 1999; Kline

2024, 146): the variance of the Grade 8 grade 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 grades 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 25: **The wedge explains much of the Grade 8 grade and barely covaries with destinations**

	2018		2022	
Share of the variance (%)	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 × Cov(persistence, wedge)	−16.2	−16.2	−9.8	−9.8
2 × Cov(persistence, destination)	1.5	1.5	−1.0	−1.0
2 × Cov(wedge, destination)	3.2	4.0	−3.0	−2.3
Residual	31.5	32.2	38.5	39.1
Variance of the Grade 8 grade (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 grade, 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. Grades 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: outputs/tables/mathematics/tab_akm_appendix_decomposition.csv, built from analyses/wedge_noise_correction/outputs/tables/variance_decomposition.csv and variance_decomposition_validation.csv; each number's origin is listed 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).

I.4: 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 grades below that joinable group. The sign of the gap is an artefact of the cut-off, not a fact about selection. Figure 6 plots it as the cut-off moves from the whole attrited pool to the 30th percentile: linked students sit 5.5 grades above the untruncated pool, the gap crosses zero close to the tenth percentile, where the two groups differ by 0.45 grades, and reaches 5.1 grades below by the thirtieth, because each higher cut-off drops more low-scoring non-progressors from the comparison group and lifts its mean.

Table 26: 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 *d*	*p*-value
Mean Grade 7 grade	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
% 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 grade, 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 (grade 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.*

I.5: 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 27: 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 (points)
Mean school effect (IPW-weighted)	0.22	IPW-weighted mean school effect (points)
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,

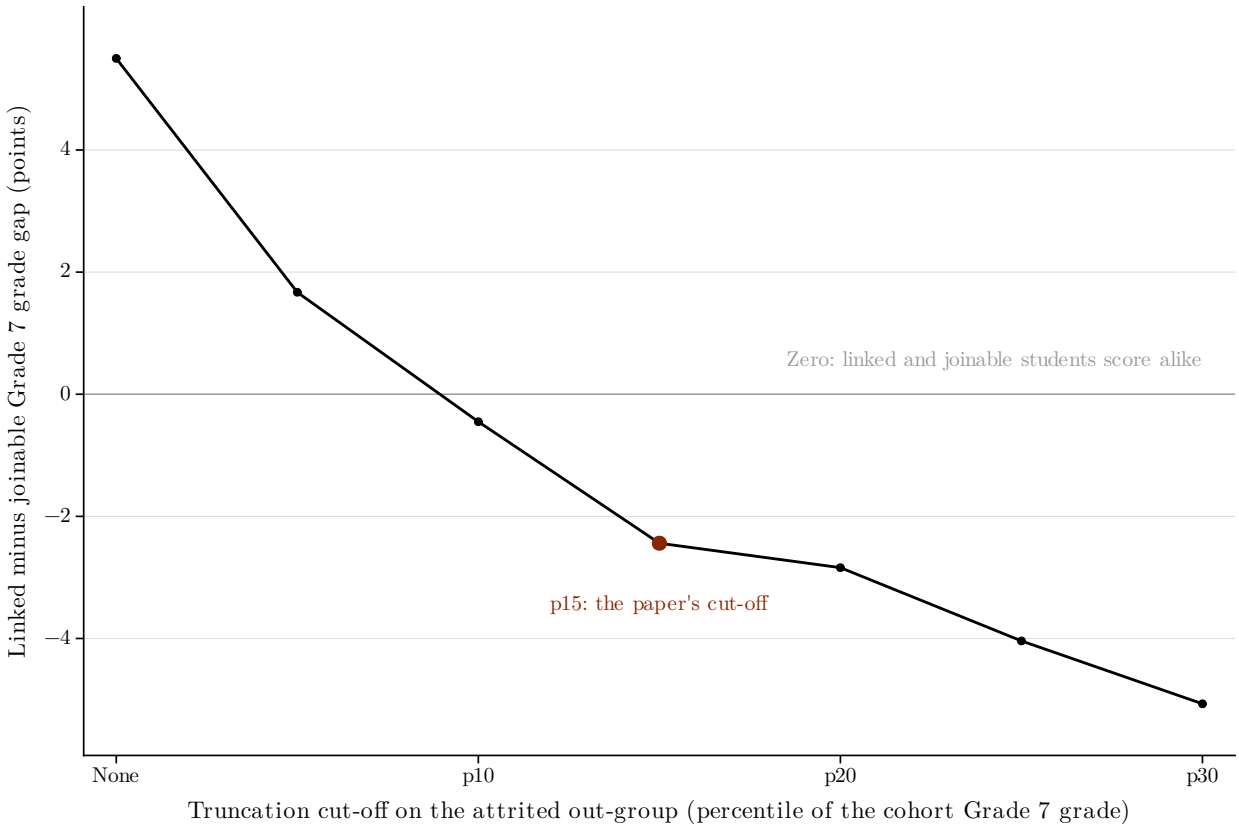

Figure 6: **Whether linked students score above or below the joinable out-group depends on where that out-group is truncated.** *Linked-minus-joinable Grade-7 grade gap as the truncation cut-off moves from none (the whole attrited pool) to the 30th percentile of the cohort baseline. A higher cut-off 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.*

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.

Appendix J: Additivity: one shift per school

The model gives each primary school one shift, the same for every student. If a school graded some of its students more leniently than others, one shift would misstate its wedge. This is the largest departure from the model: the size of a school's wedge depends on which of its students one looks at, but its rank changes little when each Grade 7 quintile gets its own slope and shift. Net of the Grade 7 margin every feeder shares, low-grade students at the feeders with the easiest assessments land 0.324 SD above the model's prediction in 2018, against a wedge standard deviation of 0.453, yet a slope and shift for each Grade 7 quintile leaves the order of feeders almost unchanged (Appendix J.1). Wedges estimated on a school's over-age and on-age students, or on its girls and boys, order schools alike (Appendix J.2), and other functional forms recover the same order (Appendix J.3). The records cannot tell classes apart, so a school's wedge is the average over its classes, and differences between Grade 8 teachers count as error. Nor do they record household wealth, and too few schools teach enough students of two population groups to compare them (Appendix J.2).

J.1: 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 28 and Figure 7 report both checks for Gauteng.

Table 28: **The two school effects nearly add; quintile slopes leave the order of feeders almost intact**

Statistic	2018	2022
Panel A. Pairs of feeder and destination		
Adjusted R^2, one effect per feeder and per destination cohort	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 mean residual, wedge decile × destination decile (SD)	0.039	0.037
Joint test that the mean residuals are zero, p-value	0.020	0.194
Panel B. One shift per school		
Largest absolute mean residual, Grade 7 decile × wedge decile, net of the Grade 7 margin (SD)	0.369	0.374
Largest absolute mean residual, Grade 7 decile × destination decile, net of the Grade 7 margin (SD)	0.248	0.105
Panel C. A slope and shift per Grade 7 quintile, shared by all feeders		
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 cohort, 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 Sølvsten (2020) on the links that carry two or more students, and reports the largest absolute mean residual over the pairs of wedge decile and 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 × wedge decile, and by Grade 7 decile × destination decile, each net of the Grade 7 margin shared by all schools: each decile pair'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 grade. Pairs of deciles 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 grades. Not run here: the change in grades 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: 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; each number's origin is listed 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 to the adjusted R^2, and mean residuals by wedge decile and destination decile are small (panels A and B of Figure 7), but these feeder–destination interactions are jointly significant in 2018, though not in 2022 (Table 28). 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 grade cannot be measured. Over all links the SD therefore rests on an assumption about it: 0.134 in 2018 if those students' grades 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 grade 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).

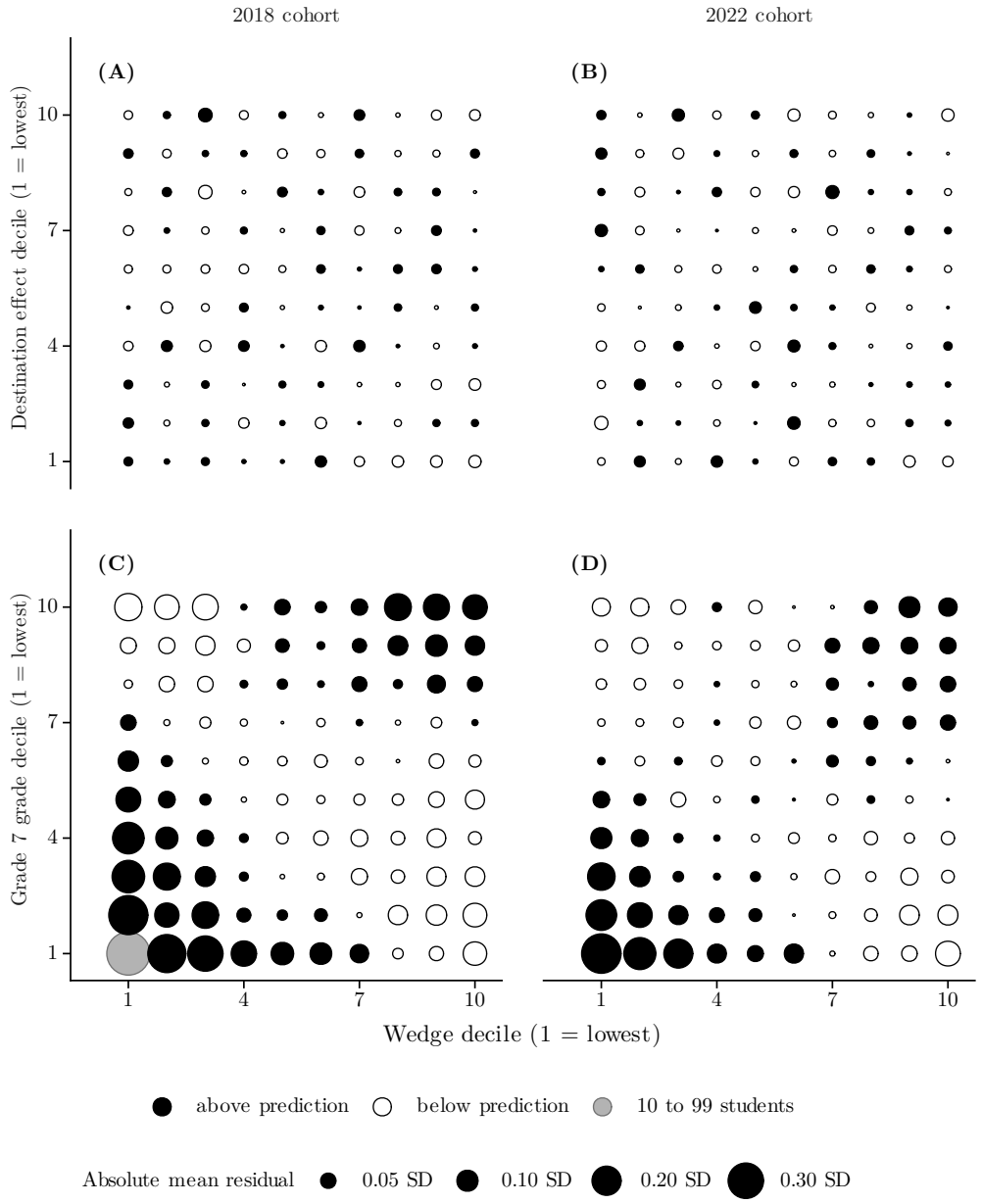

Figure 7: **Residuals are small across pairs of schools but vary with student grades within a feeder**

***Notes:** Each circle is the mean residual of the model of the main text for one pair of deciles, in within-province standard deviations (SD) of the grade: its area is proportional to the absolute mean, filled where the pair's students score above the model's prediction and open where they score below. Under the model, with each school one shift, every pair'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 grade, net of the Grade 7 margin shared by all feeders: each pair'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 grades. Grey circles are pairs of deciles holding 10 to 99 students (1 in all); pairs with 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_za.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; pairs of deciles holding 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 grade (panels C and D): low-grade students of feeders with easier assessments and high-grade students of feeders with harder ones score above the model's prediction, the opposite pairs below it. Net of the Grade 7 margin, low-grade 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 28). A slope for each feeder takes up most of the pattern: the largest net residual 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 grade, but a feeder's wedge for students at the 10th percentile of the Grade 7 grade 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 grades relative to its students' Grade 8 grades. 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 Grade 7 and Grade 8 grades, and most of it disappears once the term-to-term reliability of the feeder's Grade 7 grades and the spread of its students' Grade 8 grades are held fixed. At the destination, the Grade 8 grade rises more steeply with the Grade 7 grade 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 B.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 grade dependence is common to every feeder. Among students who leave the same primary for the same secondary school, the Grade 8 grade bends away from a single line fitted within each pair of schools at both ends of the Grade 7 range: in Gauteng, 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 grade within two points of 30%, 40% or 50%, and every zero, leaves it at least as wide. Nor does a floor in the Grade 8 grade: 16 per cent of students below 30% in Grade 7 score below 10% in Grade 8. A single line through two term grades 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 grades 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.

J.2: 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 8 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.

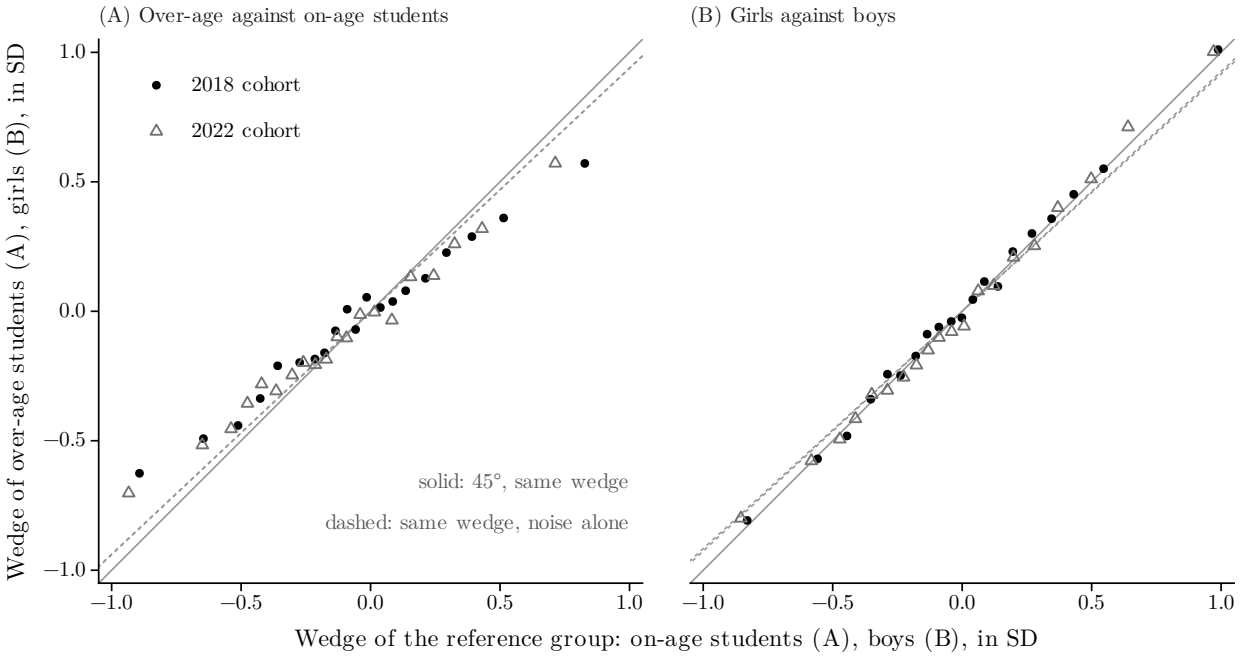

Figure 8: **Over-age students' wedges are narrower by their flatter Grade 7 slope; girls' slightly wider**

***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 grade. 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 grade by the same amount lowers each group's wedge by that amount times the group's own slope, so the group whose grades 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 grades 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 grades are not set on a narrower scale. Their lower grades explain part of the flatter slope: a slope per feeder, read at each group's mean grade, would give the over-age wedge a slope of 0.87 and 0.95 on the on-age wedge (Appendix J.1). 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 grades track their Grade 8 grades 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 grades 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 J.1). 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 grade. 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.

J.3: Non-parametric estimation alternatives

In Gauteng, 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 29, 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 grades within their Grade 8 school. It does not net out the Grade 7 grade, 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 grade 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 → Grade 8 curve to grades 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 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 29: 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 (points)	Median SE (points)
Baseline TWFE	Linear, OLS	1.000	9.02	1.251
Quadratic TWFE	Add Gr7² as additional regressor	0.994	8.52	1.246
Piecewise linear	Hockey-stick spline at median Gr7 grade	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 grade 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, 2018 and 2022 cohorts.

For downstream applications using the calibrated learning measure as a grade-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 grade-point estimators.

Appendix K: Heaping at the promotion threshold

Grades bunch at the 40% promotion threshold, because students below it are pushed up to it and never down, and the push is heaviest at the weakest schools. If the pushing reached the wedge, the order of schools would partly reflect where a school's students sat relative to 40. It does not: smoothing both grade distributions and re-estimating leaves raw and de-heaped wedges correlated at 0.9973 or above in every province and cohort, and moves the Grade 7 slope by at most 2.1 per cent (Appendix K.2, Appendix K.3). The appendix also shows heaping by school quintile and the Grade 12 reversal that ties it to whoever controls promotion (Appendix K.1), and two simulations: one that heaps every school alike at a strength set to match the observed heap (Appendix K.4), and one that heaps the weakest schools hardest (Appendix K.5). In both, the estimator orders schools as well from the heaped grades as from the clean ones.

K.1: Heaping heterogeneity across school quintiles

Table 30: **Grades heap at 40%, the promotion requirement in mathematics, and hardly at 30% or 50%.**

Grade	Heaping index	Excess mass (%)
2018 cohort		
30	1.107	9.7
40	1.935	126.0
50	1.011	4.6
2022 cohort		
30	1.062	3.5
40	2.240	146.5
50	1.046	5.5

Notes: Each row is one threshold in one cohort's Grade 7 Term 4 mathematics grades. The heaping index is the number of students at that grade divided by the mean number at the two adjacent grades; an index of one means no heap. Excess mass is the percentage by which the count at that grade exceeds a counterfactual density, smoothed by local polynomial through the grades that are not multiples of five. Promotion in mathematics requires 40%, not 30% or 50%. Sample: National (6 provinces), 2018 and 2022 cohorts.

Ranked by the DBE poverty quintile, a school-funding classification rather than a measure of academic standing, the 40% 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 grade heap about 5.2 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 K.5).

Grades heap more in Term 4, which decides promotion, than in Term 1, and the heap disappears at Grade 12, where promotion runs through the National Senior Certificate rather than through the school's own assessment (Figure 10). There is no spike at 40 in either term at Grade 12: the count at 40 is within 5 per cent of the smooth distribution around it, against about two to three times it in Grade 7. Grades 10 and 11 sit at 1.041 and 1.065; the claim rests on the Grade 12 reversal alone.

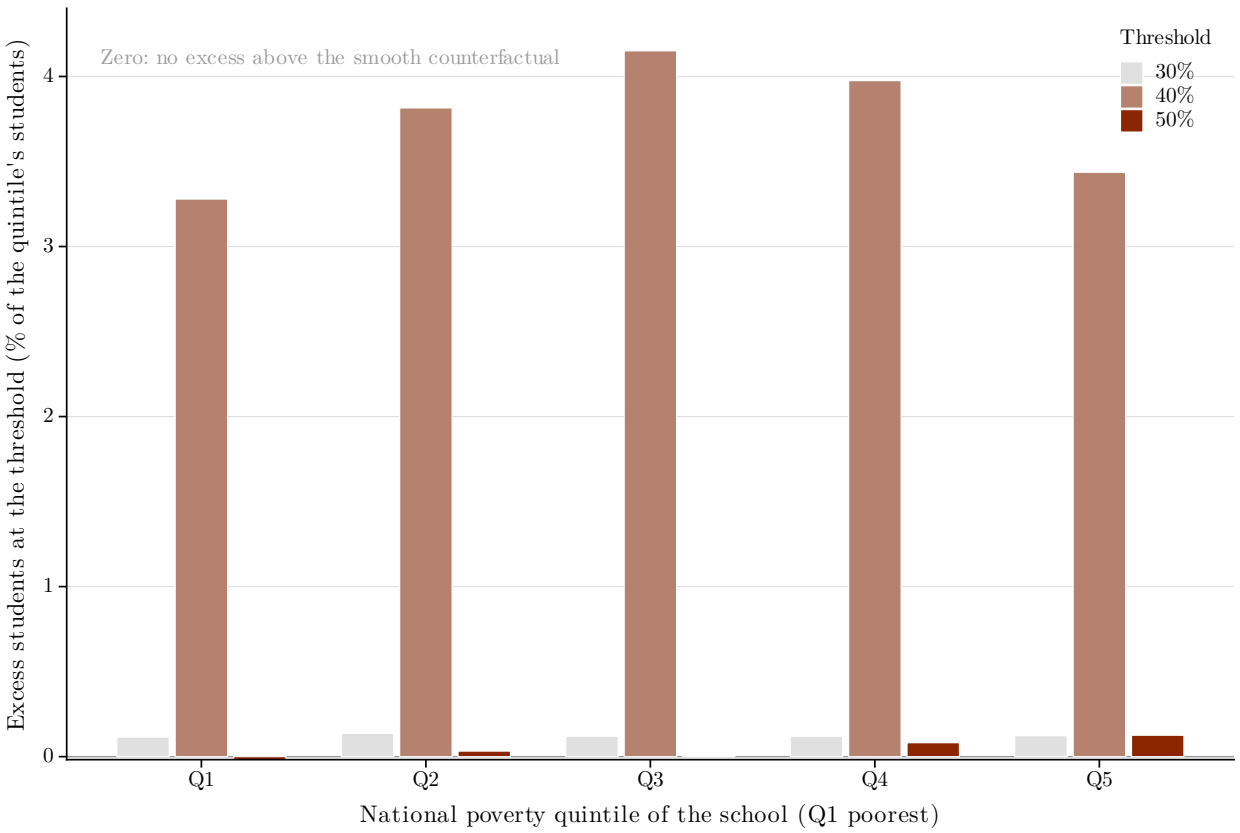

Figure 9: Excess students above the local-polynomial counterfactual at each threshold, expressed as a percentage of each DBE poverty quintile total (Q1 poorest to Q5 wealthiest, a funding classification). The 40% threshold dominates across all five poverty quintiles; the 30% and 50% thresholds contribute negligibly. Ranked by academic standing instead, heaping is heaviest at the lowest-scoring schools (Appendix K.5). *Sample:* National (6 provinces), pooled 2018 and 2022 cohorts.

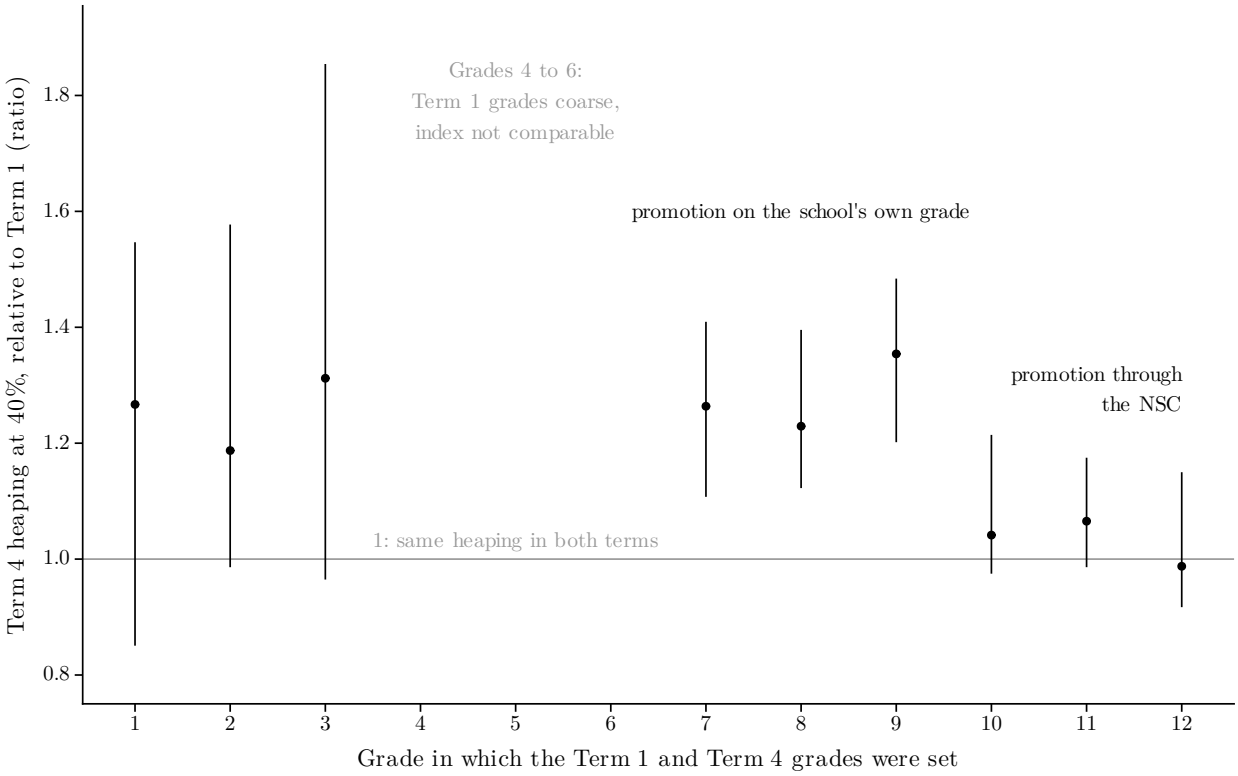

Figure 10: **Heaping tracks the term that decides promotion, and dies at Grade 12.**

Note. Each point is the median, across province-years of that grade, of the ratio between the Term 4 and the Term 1 heaping index at the 40 grade; the whiskers are the interquartile range across those province-years. The heaping index is the frequency of the 40 grade over the mean frequency of 39 and 41, so a ratio above one means the grade was pushed to the threshold more often in Term 4 than in Term 1. The grey horizontal line at one is the benchmark, drawn behind the data. Medians run from 1.187 to 1.354 through Grade 9, sit at 1.041 and 1.065 at Grades 10 and 11, and reach 0.988 at Grade 12, the grade whose promotion runs through the National Senior Certificate rather than through the school's own assessment. **The claim here is the Grade 12 reversal, not a monotone gradient across grades: the Grades 10 and 11 figures are reported and not explained.** The National Senior Certificate score is the final blended score, 25% school-based assessment plus 75% external. A province-year enters only where both terms use the full integer grade scale, and a grade only where at least ten such province-years survive; Grades 4 to 6 fail that rule and are drawn as a gap, labelled in the panel. No standard error is available, because the heaping table records a count and an index rather than a sampling variance, so no significance star appears here and none applies; 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/sba_nsc_link/outputs/tables/tab2_heaping_t1_t4_ratio.csv, via outputs/tables/mathematics/tab_nsc_heaping_by_grade.csv. **Sample:** Mathematics grades in the national school-based-assessment extract, Term 1 against Term 4 of the same year at the same school grade; 186 province-years across the 9 plotted grades, North West 2022 excluded because its Term 1 and Term 4 grades are duplicates.

K.2: Structural de-heaping: $\Delta\beta_1$ across 12 province-cohort samples

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 grade distributions, separately for each of the 12 province-cohort samples. The maximum absolute shift across all samples is $|\Delta\beta_1| = 0.0154$ (2.1% of the worst sample's baseline β_1, occurring in Eastern Cape 2018); the direction is negative in 12 of 12 samples. The de-heaping procedure fits a degree-7 polynomial counterfactual density excluding a $\pm$2-point window around 30%, 40%, and 50%, separately for each grade and for each province, cohort and quintile, and draws $m = 25$ imputed grade panels with both grades redrawn where heaped. Rubin's combining rules pool within-imputation analytical SEs with between-imputation variance.

Table 31: Structural de-heaping: $\Delta\beta_1$ by province and cohort (Rubin-pooled, $m = 25$ imputations).

Province	Cohort	$\hat{\beta}_1$ (raw)	$\Delta\beta_1$	$\lvert\Delta\rvert/\beta_1$ (%)	$\lvert\Delta\rvert$ / 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
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 grades de-heaped jointly. Sign-uniform: $\Delta\beta_1 < 0$ in all 12 of 12 province-cohort samples. Economic shift $|\Delta\beta_1|/\beta_1 \leq 2.1\%$ in every sample.

K.3: Structural de-heaping: $\hat{\alpha}_j$ Pearson ρ across 12 province-cohort samples

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

Table 32: Structural de-heaping: $\hat{\alpha}_j$ Pearson ρ, raw vs de-heaped school effects.

Province	Cohort	Schools	ρ (min)	ρ (median)	Mean shift (points)	SD shift (points)
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 and cohort. 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.

Table 33: **The choice of measurement-error correction does not change the school-covariate gradients.**

Covariate	(A) Raw	(B) Baseline	(C) Multi-term	(D) De-heaped
Quintile 2	-0.288 (0.152)	0.346 (0.152)	0.332 (0.150)	0.353 (0.151)
Quintile 3	-0.439 (0.166)	0.562 (0.167)	0.608 (0.165)	0.557 (0.166)
Quintile 4	0.356 (0.278)	4.251 (0.280)	4.286 (0.276)	4.229 (0.278)
Quintile 5	2.990 (0.310)	8.273 (0.311)	8.400 (0.308)	8.227 (0.309)
Log enrolment	-0.964 (0.111)	1.655 (0.111)	1.450 (0.110)	1.644 (0.110)
Pupil-teacher ratio	-0.007 (0.012)	-0.158 (0.012)	-0.150 (0.012)	-0.156 (0.012)
Over-age rate	-13.674 (0.472)	3.941 (0.474)	4.274 (0.469)	3.921 (0.471)
R^2	0.094	0.121	0.120	0.121
RMSE	6.581	6.606	6.534	6.562

Notes: Each row is a school characteristic and each column a regression of a school-level outcome on them. Column A: the raw Grade 7 school mean, in grades. Columns B–D: the wedge alone, in Grade 8 grades, under the baseline, multi-term and de-heaped measures respectively. Column A and columns B–D are different outcomes in different units, so compare B–D with one another, not with A. The wedges are estimated in each province on its own; each regression then stacks the six provinces' schools with a fixed effect for each province, and R^2 is the adjusted within-province R^2. Quintile 2 to Quintile 5 are the school's poverty quintile, each against the poorest, Quintile 1. Sample: National (6 provinces), pooled 2018 and 2022 cohorts.

K.4: Monte Carlo recovery under threshold-pushing

The figure below reports the heaping simulation described in §5.3 of the main text. Panel A shows the baseline coefficient $\hat{\beta}_1$ under three estimators, namely oracle (computed on pre-heaping grades), OLS with heaped grades, 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 grades in Figure 12 confirm that threshold-pushing is active in the DGP, set to a target heaping index of 2.4 at the 40% threshold. The spike at 40 accounts for 2.7% of heaped grades against 1.6% in the smooth reference.

K.5: Heaping–quality correlation bounding

The heaping simulation of §5.3 of the main text heaps 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 Gauteng's schools the per-school heaping share correlates with the raw mean Grade-7 grade 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 grade are themselves negatively correlated. A school low on the raw grade but high on value-added heaps most. Since heaping only raises grades, 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 two school traits, run separately: decreasing in the raw mean Grade-7 grade, 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 gradient, so the sweep isolates the correlation from the severity that Appendix K.4 already bounds. On the raw-grade axis the gradient is calibrated to Gauteng's data and pushed past it, to a bottom-to-top quintile ratio of 9 against the data's 5.6; on the school-effect axis it never reaches the link between heaping and the wedge that

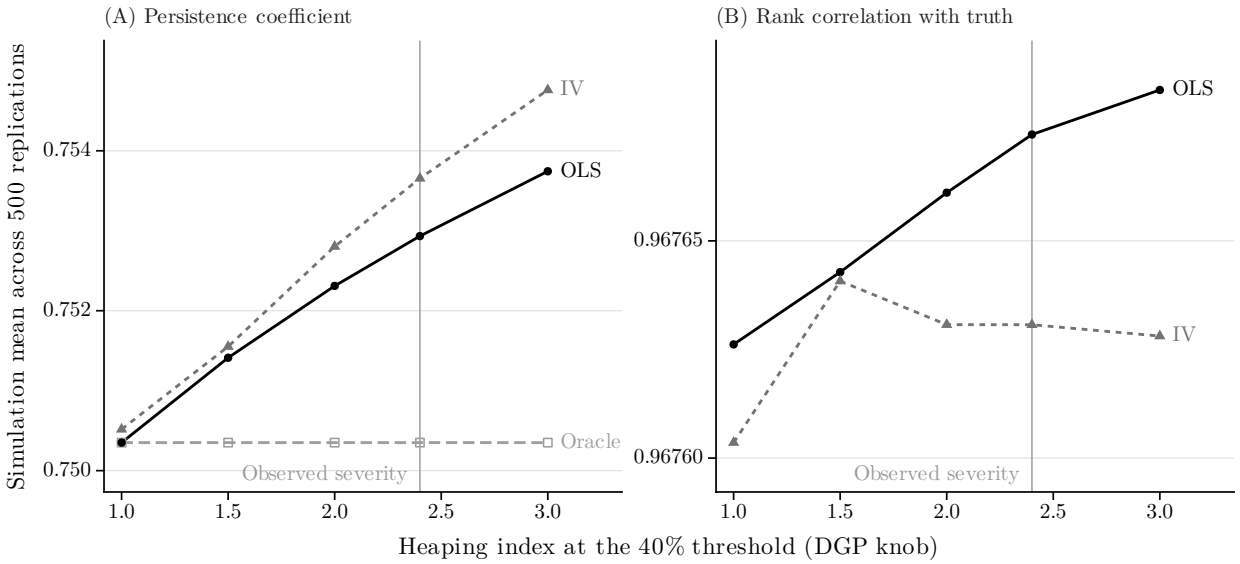

Figure 11: Monte Carlo: threshold-pushing in Grade 7 grades. Panel A: baseline coefficient under oracle, OLS, and IV. Panel B: Spearman rank correlation between estimated and true school quality. The grey vertical line marks the observed heaping index. 500 replications per scenario. *Sample:* simulated data calibrated to Gauteng, 2018 and 2022 cohorts.

the data show (§5.3 of the main text). Across the sweep the rank correlation with the truth stays within 0.0004 of the clean-grade oracle, the Grade 7 slope moves by about 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 quintile of true school effects, where the inflation would concentrate, stays under 0.10 of a point against $\sigma_\tau \approx 6$. Directional heaping tied to the school's mean grade, at the gradient the grades show and steeper, does not reorder schools (300 replications per setting).

Appendix L: Inference with an estimated wedge

Two of the paper's applied results use the wedge in a second regression: the production function, and the atlas of which schools set easier assessments. Each could overstate what it finds. The wedge is estimated, so a second stage that treats it as known understates the uncertainty, and a search across many school characteristics will find some by chance. Neither changes the results. Redrawing every wedge in a paired cluster bootstrap leaves the production function's intervals not substantially larger and changes the significance of no coefficient (Appendix L.1). In Gauteng, after a Romano–Wolf adjustment for the number of characteristics tested in each family, the no-fee indicator, at an adjusted p below 0.001, and the school's own poverty quintile, at 0.008, still separate schools with easier assessments from the rest (Appendix L.4). Appendix L.1 also sets out how to carry the first-stage error into other work, and Appendix L.2 to Appendix L.5 hold the atlas: its variables, every coefficient, and the split of the wedge's variance among families of characteristics.

L.1: Using the estimates downstream

This subsection 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 $\widehat{\mathrm{se}}(\hat{\alpha}_s)$, with one file per province and a pooled file across provinces. Both

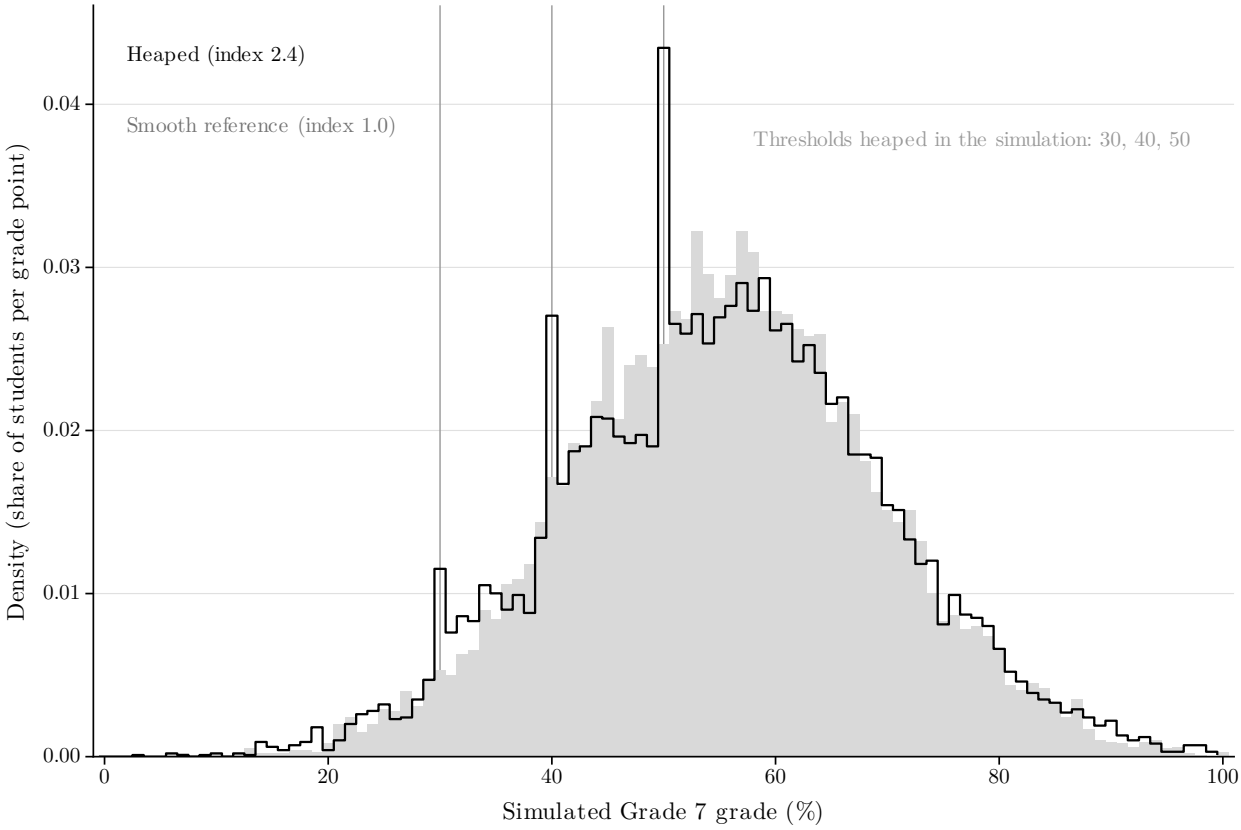

Figure 12: DGP face-validity: threshold-pushing in generated Grade 7 grades. Density of heaped grades (DGP setting: heaping index 2.4 at the 40% threshold) overlaid on the smooth unheaped truth, with vertical lines at the 30, 40, and 50 thresholds. The spike at 40 is visible: 2.7% of heaped grades land exactly at 40, against 1.6% in the smooth reference (a 1.7-fold excess). The empirical Gr7-2022 heaping index is 2.24. *Sample:* generated data from the heaping simulation of §5.3 of the main text, not observed students.

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 grades), `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 9 implies that in Gauteng this ratio is small, but the correction is not negligible and is larger in provinces with sparser mixing networks (Table 6).

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 and can be substituted when the application tolerates the shrinkage-induced bias in exchange for lower variance.

The production function of §4.4 of the main text uses the first correction, and Figure 13 sets its bootstrap intervals beside the naive ones.

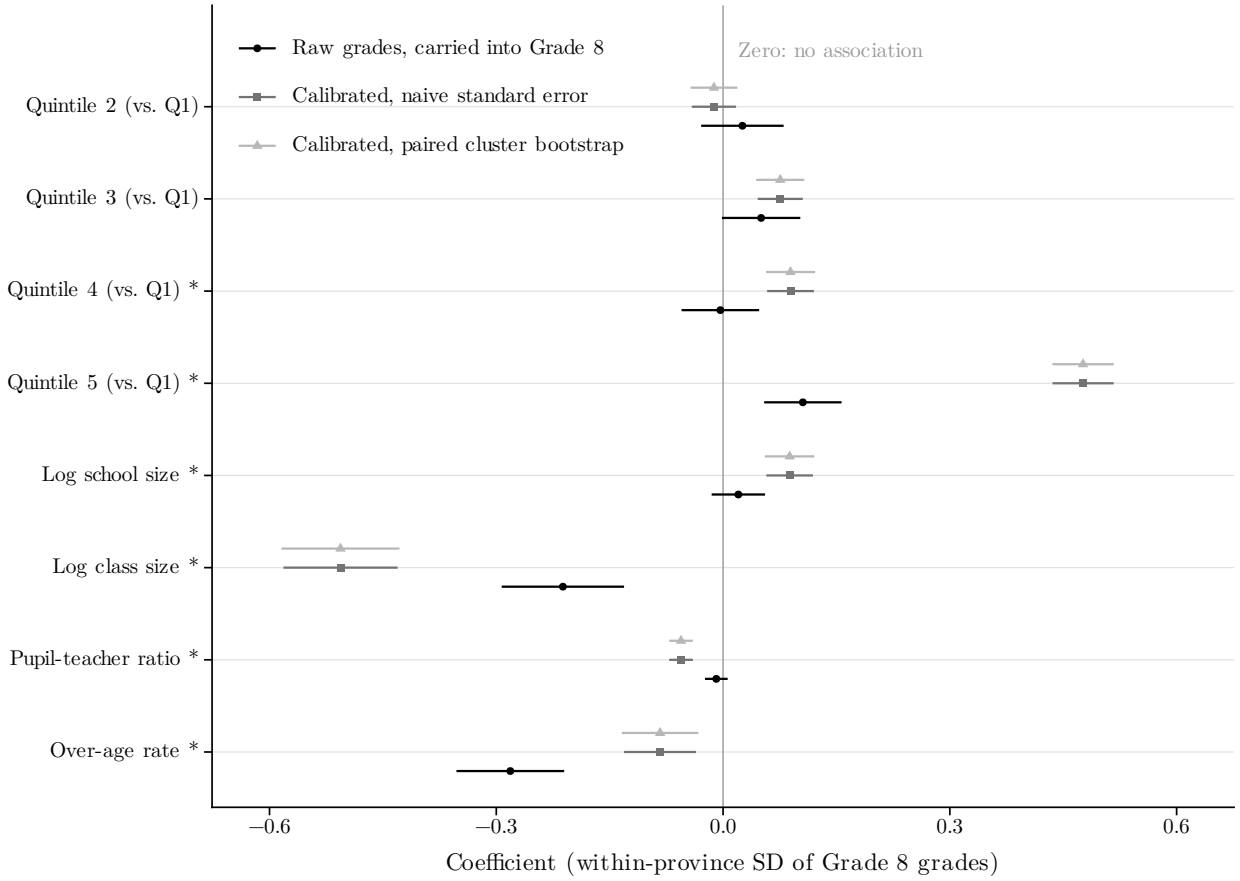

Figure 13: **Calibration moves the poverty gradient most; the bootstrap barely widens the intervals.**

Notes: Rows are coefficients from a regression of the school's raw mean Grade 7 grade, or its calibrated learning level, on poverty quintile (against Quintile 1), log enrolment (log school size), log class size, the student–teacher ratio (pupil–teacher ratio) and over-age, mean years above the expected age for Grade 7, with cohort effects; the gradient in §4.4 of the main text leaves over-age out. Raw coefficients are multiplied by persistence in their province and cohort, so both series are in within-province SD of Grade 8 grades. Bars are 90 per cent intervals; a star marks a shift whose 95 per cent bootstrap interval excludes zero. Sample: Gauteng, 2018 and 2022 cohorts, 2,332 school-cohorts in 1,283 primary schools.

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 C). 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 F.2). 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.

L.2: The covariate atlas

This subsection and the three after it reproduce Gauteng's covariate-atlas exhibits, which support §4.5 of the main text. All exhibits are re-formats of existing wedge-supply outputs; no new estimation is performed here.

L.3: Covariate atlas, variable dictionary

Table 34: Covariate atlas — variable dictionary. Every regressor plotted in Figure 9 of the main text and listed in Table 35, followed by the mechanical controls held fixed in each regression.

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

(continued)

Variable	Family	Definition
`gender_ratio_f`	Student composition and churn	Share of girls among the school's linked Grade 7 Mathematics students in the same cohort.
`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/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 9 of the main text) and Table 35 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 36 use the standardised persistent component (`wedge_persistent`). Every regression weights school-cohorts by 1/wedge_se², 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.

L.4: Covariate atlas, full coefficient table

Table 35: Full covariate atlas: associations with the school-level grade–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	Grading style	-0.265	0.018	< 0.001	0.000	No
heap_40	Grading style	-0.222	0.018	< 0.001	0.000	No
heap_50	Grading style	-0.021	0.022	0.344	0.005	No
term4_lift	Grading style	0.332	0.030	< 0.001	0.000	No
round_number_share	Grading style	-0.349	0.038	< 0.001	0.000	No
within_school_sd	Grading 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

(continued)

Covariate	Family	Coefficient	SE	p	RW p	Artefact
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
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_secondary	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 34) marks artefact-suspect. Sample: GT pooled cohorts (2,903 school-cohorts; 1,280 GT schools with both cohorts matched). Precision-weighted = 1/wedge_se². 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.

L.5: Shapley variance decomposition

Table 36: 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 $R^2 = 0.371$. Shares sum to the joint R^2; no causal interpretation.

Appendix M: Validation against the National Senior Certificate

If calibration only reshuffled grades, it would predict a later, external result no better than raw grades do. The National Senior Certificate (NSC), the school-leaving examination, is that result for one cohort: students in Grade 7 in 2020, who sat it in 2025. Among students who entered the same secondary school with the same raw Grade 7 grade, those from a primary school with harder assessments do better: a wedge one SD higher raises the probability of sitting the NSC by 0.340, against 0.196 for an own Grade 7 grade one SD higher, and predicts the NSC score itself (Appendix M.1). That subsection gives the specifications in full, the results by province and how far the observed ratios fall short of the calibration's benchmark; Appendix M.2 gives the school-level test on the paper's own cohorts. In the national extract, school characteristics come from the 2025 masterlist, even where they are attached to a school observed in an earlier year.

M.1: The wedge against the student's own NSC score

Table 37 reports in full the specifications drawn in Figure 7 of the main text: the score 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-grade coefficient against its structural prediction, 1.91. Table 38 reports the wedge coefficient by province of the secondary school of entry, with and without the origin school's mean grade, each against that province's own prediction. The wedge is positive and significant in every province, in both subjects; the North West Mathematics estimate 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 §4.3 of the main text, a different object from the school-level rank correlation of Appendix M.2.

The conditional increment and the school-level rank correlations of Appendix M.2 are consistent, because they measure different objects. The school-level test sets a school's calibrated level against that school's NSC aggregate across schools with different intakes, while the test here holds the secondary school of entry fixed and sets the origin school's wedge against the student's own grade.

Adding the origin school's own mean Grade 7 grade as a further control pulls the wedge's coefficient down to 0.401 and 0.454. The construction predicts this. By the bridge identity of §3.4 of the main text, the wedge is roughly its students' mean Grade 8 grade minus a fraction of their mean Grade 7 grade, so controlling for the Grade 7 mean controls for part of the wedge itself, in effect racing a proxy for assessment difficulty against the measure of it: a school whose students all score high on its own assessments is one with easy assessments. That specification (M3) is therefore a diagnostic. The coefficients to read are those of M1 and on the participation margins, the same reading Appendix N.9 gives the sorting test.

Two limits bind the score regressions of Table 37. 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. The whole entering cohort is not selected on the wedge, which is why the participation margin leads. Selection of that kind would be expected to pull the score 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 score, 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.

Measurement error in the wedge accounts for most of the participation shortfall but for the smaller part of the Mathematics and Mathematical Literacy shortfalls. Once the secondary-school fixed effect is in, only within-secondary-school variation identifies the wedge's coefficient, and there its school-effect reliability, the Kane–Staiger signal share of §4.2 of the main text with the variance taken within secondary school of entry, is 0.928, against 0.951 with the variance taken over all students in the sample. Correcting for that attenuation lifts the participation ratio to 1.87 but the score ratios only to 1.64 and 1.74. The rest of the participation shortfall is the part of the 2020 wedge that does not persist across cohorts. In Mathematics, instrumenting with the 2018 and 2022 wedges moves the ratio further from the benchmark, to 1.32. The shortfall is no smaller in Gauteng, where the Mathematics ratio reaches 0.71 of the value its own calibration predicts, against 0.80 pooled; with no standard error on either ratio, I cannot test the difference.

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 grades 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 grade shares that grade'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 score runs from 1.47 through 1.57 to 1.30 (0.051), and the Mathematical Literacy score from 1.65 through 1.77 to 1.74: within the 2020 cohort the two corrections together close less than half of either score'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 grades 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 score 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 grade 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.

The Grade 8 grade gives the mirror image of the Grade 7 correlations in §4.3 of the main text. Its calibrated version is the Grade 8 grade minus the Grade 8 school's destination effect γ_k, and removing that effect raises the correlation from 0.29 to 0.36. In Mathematical Literacy the Grade 7 correlation rises from 0.25 to 0.40 and the Grade 8 one from 0.22 to 0.29. Of the 14 fits at each grade, one for each subject on the pooled sample and in each of six provinces, the calibrated grade correlates more strongly in 14 at Grade 7 and in 14 at Grade 8.

Table 37: **Students from origin schools with harder assessments sit the NSC more often and score higher.**

	Coefficient on					
	Wedge	Grade 7 grade	Calibrated score	Ratio	Within R^2	Observations
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. NSC 2025 score, students who sat						
Mathematics, M0		0.338			0.152	70,790
		(0.009)				
Mathematics, M4			0.457		0.191	70,790
			(0.005)			
Mathematics, M1	0.697***	0.458		1.524	0.195	70,790
	(0.016)	(0.005)				
Mathematics, M3	0.401***	0.532			0.208	70,790
	(0.015)	(0.005)				
Mathematical Literacy, M0		0.309			0.160	85,454
		(0.007)				

(continued)

	(1)	(2)	(3)	(4)	(5)	(6)
Mathematical Literacy, M4			0.469		0.204	85,454
			(0.005)			
Mathematical Literacy, M1	0.751***	0.463		1.620	0.207	85,454
	(0.013)	(0.005)				
Mathematical Literacy, M3	0.454***	0.516			0.214	85,454
	(0.016)	(0.005)				
Panel B. Sitting the NSC, the whole entering cohort						
Sat the NSC	0.340	0.196		1.737	—	394,353
	(0.003)	(0.001)				
Sat Mathematics	0.286	0.176		1.629	—	394,353
	(0.003)	(0.001)				
Sat Mathematical Literacy	0.054	0.020		2.678	—	394,353
	(0.003)	(0.001)				
Mathematics, given sitting	0.365	0.228		1.602	—	156,244
	(0.005)	(0.002)				

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 score, standardised nationally over every 2025 sitter of the subject. The NSC score is the final blended score: 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 grade, read from analyses/hs_value_added and never refitted; the own Grade 7 grade is the student's Grade 7 Term 4 grade, standardised within province and subject on the calibration sample, so it is in standard deviations of the Grade 7 grade, 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 grade and the controls; M1 adds the wedge; M3 adds the origin school's own mean Grade 7 grade as well. M4 is not a step of that ladder. It replaces the own Grade 7 grade with the calibrated score, the own grade and the wedge combined at the weights the calibration fixes and standardised within province and subject on the calibration sample, as the own grade 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 grade 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 score 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 main 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 grade, 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

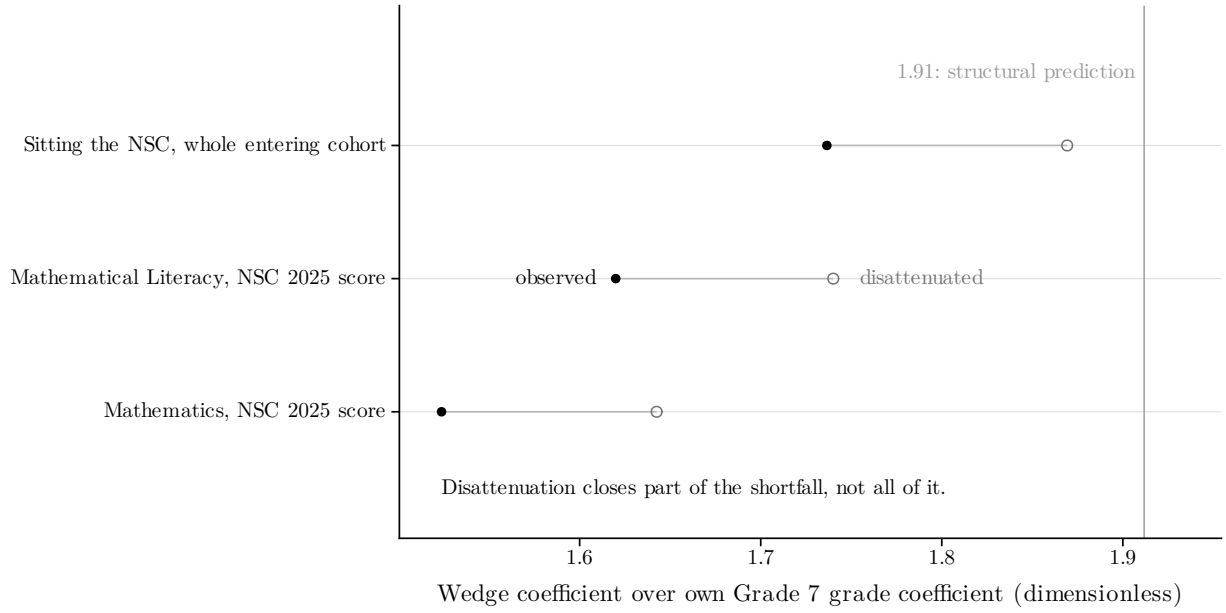

Figure 14: **Every plotted ratio falls short of the structural prediction, and disattenuation closes only part.**

Note. Each row is the ratio of the wedge coefficient to the coefficient on the student's own Grade 7 grade in the M1 specification, for NSC Mathematics and for Mathematical Literacy, and on the participation margin the ratio of the two coefficients in the linear probability model for sitting the NSC at all. The structural model predicts that ratio equals the reciprocal of the calibration slope, 1.91, which is the grey vertical benchmark drawn behind the data. The lighter open marker on each row is the same ratio disattenuated for the measurement error in the wedge, joined to its observed counterpart by a thin grey segment; disattenuation moves each ratio part of the way to the benchmark and stops, so the conjecture that attenuation alone explains the shortfall does not survive. **The claim is about the three rows drawn here and does not generalise to every margin**: the sitting-Mathematical-Literacy margin, which this figure does not plot, overshoots the benchmark. The standard errors of the three observed ratios, by the delta method with errors clustered at the origin school and the wedge treated as known, are 0.015 on the participation margin, 0.035 in Mathematics and 0.026 in Mathematical Literacy; **no interval is drawn and the ordering among the rows is not tested**. The NSC score is the final blended score, 25% school-based assessment plus 75% external. No significance star appears in this figure, and none applies; 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/alpha_nsc_validation/outputs/tables/tabA2_ratios.csv, tabA2_participation.csv and tabA3_attenuation.csv, and analyses/wedge_ratio_shortfall/outputs/tables/tabW20_a30_all_margins.csv for the participation row's disattenuated ratio and the three standard errors. **Sample:** the 394,353 students of the Grade 7 2020 cohort who entered a secondary school in one of six provinces carrying a calibrated intake grade; the two score ratios are the 70,790 who sat NSC Mathematics and the 85,454 who sat Mathematical Literacy, and the participation ratio is the whole cohort.

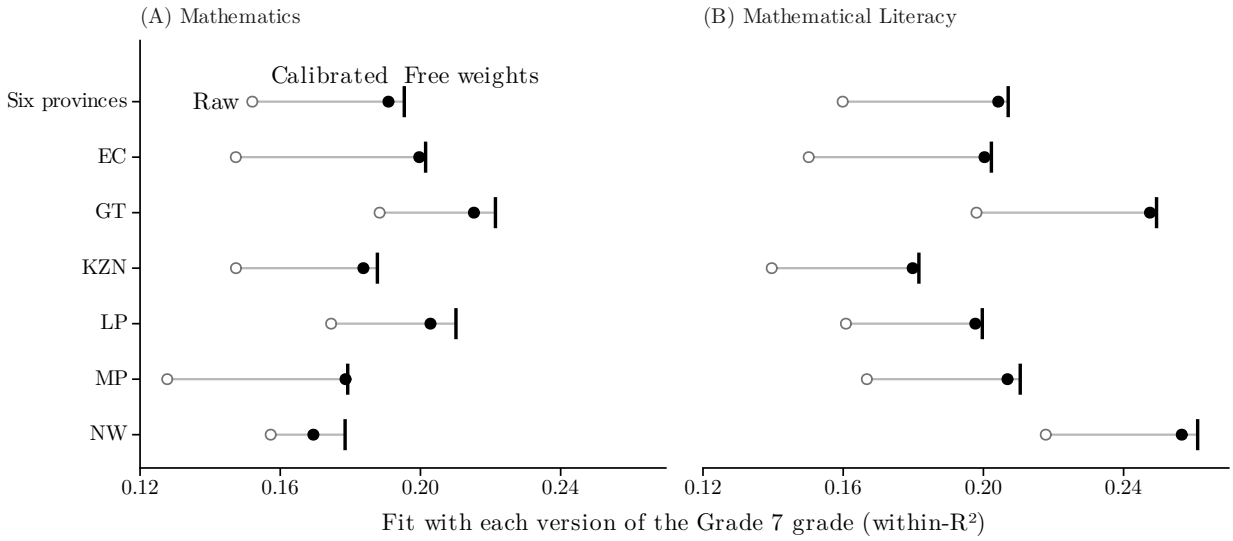

Figure 15: **The calibration's fixed weights predict nearly as well as weights the data choose freely.**

Notes: Each row is one subject in the six provinces pooled or in one province (EC Eastern Cape, GT Gauteng, KZN KwaZulu-Natal, LP Limpopo, MP Mpumalanga, NW North West). Each marker is the within-R^2, the share of the variation in NSC scores within secondary schools that the regression explains, of a regression of the student's NSC 2025 score on a version of her Grade 7 grade, with a fixed effect for the secondary school she entered and the same controls throughout: the raw grade (open circle), the calibrated score (filled circle), or the wedge and the raw grade with weights the regression chooses freely (tick). Pooled, the free weights add 0.005 in Mathematics and 0.003 in Mathematical Literacy. Sample: the Grade 7 2020 cohort in six provinces, 70,790 students who sat NSC Mathematics in 2025 and 85,454 who sat Mathematical Literacy.

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 grade. 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 38: **The wedge predicts the NSC score in every province and both subjects.**

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

(continued)

	(1)	(2)	(3)	(4)	(5)
EC	0.749***	0.431***	1.637	1.901	19,440
	(0.025)	(0.031)			
GT	0.810***	0.580***	1.551	1.815	12,794
	(0.029)	(0.041)			
KZN	0.784***	0.467***	1.796	2.113	19,906
	(0.032)	(0.039)			
LP	0.692***	0.446***	1.658	1.988	13,978
	(0.037)	(0.040)			
MP	0.679***	0.371***	1.425	1.758	11,179
	(0.031)	(0.051)			
NW	0.746***	0.442***	1.470	1.896	8,157
	(0.039)	(0.050)			

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 score, standardised nationally over every 2025 sitter of the subject. The NSC score is the final blended score: 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 grade, read from analyses/hs_value_added and never refitted; the own Grade 7 grade is the student's Grade 7 Term 4 grade, standardised within province and subject on the calibration sample, so it is in standard deviations of the Grade 7 grade, 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 grade and the controls; M3 adds the origin school's own mean Grade 7 grade 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 grade 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 sample size, which M3 shares. The North West Mathematics sample is the smallest of the exercise and gives 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 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/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. Sample sizes run from 4,575 to 19,906 students.

M.2: School-level validation against NSC results

This subsection 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 binned rank 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.

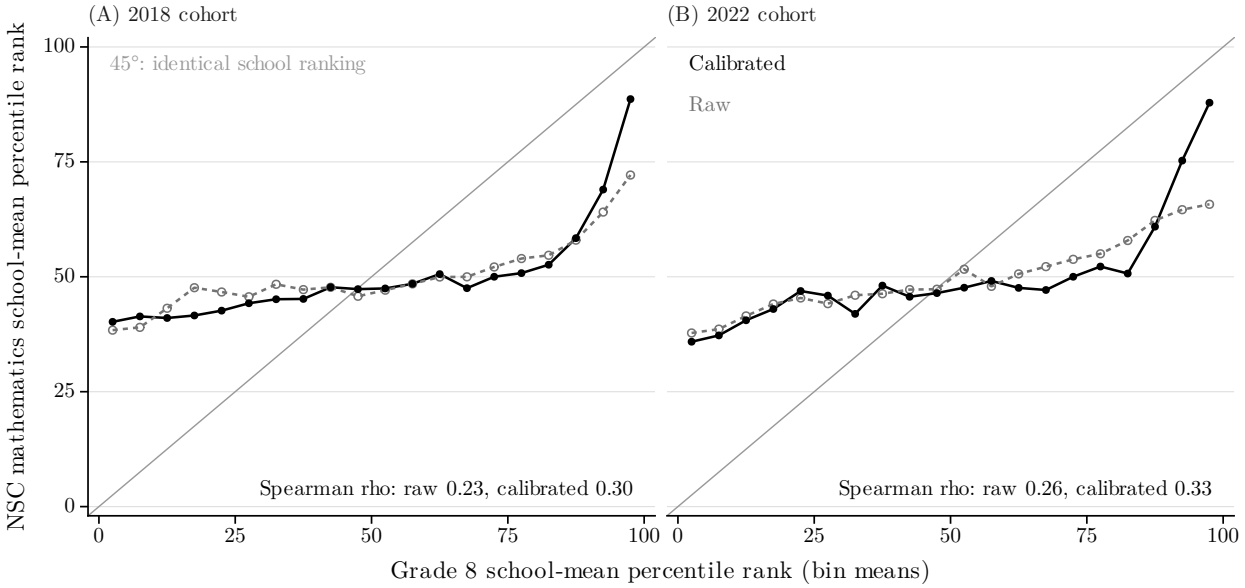

Figure 16: Binned rank plots: Grade 8 school means (raw and calibrated) against NSC 2022 Mathematics school means, both as percentile ranks. Each point is the mean of a bin of secondary schools ranked on their Grade 8 mean. Raw: the raw Grade 8 mean, in within-province standard deviations. Calibrated: the raw mean minus the school's endline fixed effect, each effect taken relative to the mean school of its province and cohort. Panel A is the 2018 cohort and Panel B the 2022 cohort; the 45-degree line marks identical rankings, and Spearman ρ is annotated in each panel. *Sample:* secondary schools of the six-province Mathematics panel matched to an NSC 2022 Mathematics school mean, by cohort.

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 grade are four years of schooling apart, and the link between a school-set grade and the examination weakens steadily with that distance. Regressing the NSC Mathematics score on the school-set Mathematics grade with a fixed effect for the school that set the assessment, 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 score, and they score 1.39 standard deviations above the rest on their Grade 8 grade.

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 grade; 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 grades its own Grade 12 students. Against the gap between a secondary school's Grade 12 Term 4 grade and its students' NSC scores, 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 grades set against the NSC correlate with its own Grade 12 grades at 0.239 (precision-weighted correlations, within province). The Grade 12 grade 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 extract's other provinces, not $\hat{\gamma}_k$.

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

Appendix N: Sorting into secondary schools

Students who leave primaries with harder assessments join Grade 8 cohorts that go on to attain more, holding fixed their own grades and the raw grades of the school they left. That could come from the feeder-mates who move with them, from where their primary school sits, or from abilities the wedge happens to track, rather than from anything about assessments. With every classmate from her own primary school left out, a student from a primary whose wedge is one SD higher joins a Grade 8 cohort that scores 0.064 SD of the peer mean higher across the six provinces (0.136 in Gauteng), while the raw Grade 7 grade predicts +0.012 (+0.020 in Gauteng) (Appendix N.3, Appendix N.10). The result holds among movers who cross a ward or municipal boundary in Gauteng (Appendix N.12), with each cohort's choices set against the other cohort's wedges (Appendix N.11), and in the destinations' published matric results (Appendix N.6, Appendix N.13). The deliberation behind the pattern is not identified, and placement rules that differ by province and year are not separated from household choice. The other subsections give the question, the design, the sample and the threats in full, for Gauteng.

N.1: The question

With no externally comparable assessment at primary level, where the calibration is identified, §4.3 of the main text tested it downstream, against students' National Senior Certificate results five years after Grade 7. This appendix tests it on a decision families make at the transition itself: where their children enrol. The calibration of §3 of the main text places thousands of differently assessed primary schools on one Grade 8 scale, and where students enrol at the Grade 7 to 8 transition tracks that same scale (Figure 12 of the main text). The alignment makes the wedge more than a statistical correction, a quantity enrolment reflects at a margin where nothing carrying it is published.

The school-choice literature has mostly measured how households respond to a quality signal published at the choice margin: test scores (Hastings and Weinstein 2008; Andrabi et al. 2017), estimated value-added (Abdulkadiroğlu et al. 2020), or school quality capitalised into house prices at attendance boundaries (Black 1999). Admission to South African public schools is open by design, a quasi-market in which households may apply across schools (Woolman and Fleisch 2006). Here no learning result is published for primary schools at all. The only published results come from the school-leaving examination, the criterion of §4.3

of the main text, and for the students making this choice they arrive years later. Whatever signal operates at this transition reaches households informally; the alignment holds as if on a quantity no household could have looked up.

The test is the complement of the within-feeder permutation test of Appendix H.2, which asked whether students from one feeder split across destinations on their Grade 7 grade and on other traits, and found that they split mainly on the grade. The estimated wedge $\hat{\alpha}_j$ is identical for every student who leaves a school, so it cannot generate within-feeder splitting and that test is silent on it. The question here is whether feeders as a whole send their students to systematically different places.

Two limits bound what the test can show. Conditional on the raw grade a school's wedge and its calibrated learning level move together, so the single-signal fits identify the scale on which students sort, calibrated or raw, and not which part of that scale they respond to; a fit holding both signals checks that each enters with the weight the calibration gives it (Table 46). And it recovers the alignment whether or not any household reads a grade in light of how hard its school's assessments are; the deliberation, if any, behind the pattern is not identified. Nor is the alignment identified against unobserved Grade 7 ability that sorts the same way as the wedge, a threat taken up in Appendix N.5.

N.2: Empirical design

Three origin-school signals could predict where a student moves, each standardised over the moving students of its cohort, so that one unit is a standard deviation of that signal rather than the within-province standard deviation of students' grades of §3 of the main text: the raw mean Grade 7 grade $\overline{y}_j^{G7}$, the raw mean Grade 8 grade $\overline{y}_j^{G8}$ that the same leavers went on to earn, and the wedge $\hat{\alpha}_j$ of §3 of the main text, positive where a school's grades sit below its students' anchored learning. The regressions use the wedge rather than the school's calibrated learning level because the level is to a first order the raw Grade 8 mean already in the comparison (§3.4 of the main text); a fit on the level would reproduce the raw Grade 8 row. The wedge is the one signal that separates sorting on the calibrated scale from sorting on raw grades.

For each signal $\mathcal{S}_j$ and each destination outcome the fit is

$$\overline{Y}^z_{k(i),-i} = \beta_{\mathcal{S}}\, \mathcal{S}^z_{j(i)} + \gamma\, y_i^{G7,z} + X_i'\delta + \lambda_{p(k)\times c} + \varepsilon_i, \tag{1}$$

with the outcome $\overline{Y}^z_{k(i),-i}$ the leave-one-out peer mean of the cohort student i joins, measured as its raw Grade 7 intake grades or its Grade 8 Term 2 grades.[2] The Grade 8 grade is earned within six months of secondary entry, so both outcomes measure the composition of the cohort a student joins rather than the receiving school's later contribution.[3] The peer mean is what reaches a household about a destination, through reputation, older siblings and neighbours, and it is what families weigh where signals do exist: facing published quality measures, they choose on the achievement of a school's peers more than on the school's own contribution (Abdulkadiroğlu et al. 2020), and publicising a value-added ranking moved neither enrolment nor fees (Mizala and Urquiola 2013).

N.3: What survives leave-own-feeder-out

A mover's destination peer mean includes the feeder-mates who moved with her, so a signal that merely grades her origin can predict the peer mean through those co-movers without any household choosing anything.

[2] i indexes a moving student, $j(i)$ her origin Grade 7 school, $k(i)$ her destination Grade 8 cohort, $p(k)$ its province and c the cohort year. $\mathcal{S}^z_{j(i)}$ is the one origin signal entered, $y_i^{G7,z}$ the student's own Grade 7 grade, X_i the controls (gender, home language, race, over-age status, origin quintile, urban-rural status, no-fee status) and $\lambda_{p(k)\times c}$ a province-by-cohort fixed effect. The z superscript marks standardisation over the moving students of each cohort, each variable by its own spread, so every coefficient is in standard deviations of the outcome per standard deviation of the signal. Standard errors are clustered two-way at the origin Grade 7 and destination Grade 8 school.

[3] Each destination school sets and grades its own Grade 8 assessment, so the Grade 8 peer mean is comparable within a secondary school but not across secondary schools. That incomparability sits in the outcome, so it widens the standard error without biasing the coefficient; it would bias the estimate only if primary schools with harder assessments fed secondary schools with systematically higher destination effects. Netting each destination's estimated effect $\hat{\gamma}_k$ out of the peer mean raises the coefficient rather than lowering it (Appendix N.5).

Leave-own-feeder-out (LOFO) drops the co-movers from the peer mean and re-estimates. Table 39 applies it to all three origin signals on both destination margins, each a separate estimate of Equation 1, and Figure 17 draws the full-sample columns.

Table 39: **Enrolment tracks the calibrated scale, not raw grades; the raw Grade 7 advantage is own-feeder leakage.** *Each cell is the coefficient (two-way clustered SE) on one standardised origin-school signal from a separate single-predictor fit, in within-cohort SD units. The outcome is the leave-one-out peer mean of the destination cohort's raw Grade 7 intake (columns 1 and 2) or its Grade 8 Term 2 grades (columns 3 and 4). Full leaves out the index student; LOFO also drops her own-feeder co-movers (n=220,973, 99.9% retained), and the wedge's LOFO significance is bootstrap-confirmed (Figure 18). The raw Grade 8 row is a mechanical benchmark: Grade 8 grades are unpublished, so no household can act on it. Controls: own Grade 7 grade, gender, home language, race, over-age, origin quintile, urban-rural, no-fee, province-by-cohort fixed effects. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.* ***Sample:*** *n = 221,093; 1,371 origin / 883 destination schools; Mathematics; Gauteng; 2018 and 2022 cohorts.*

	Grade 7 intake		Grade 8 achievement	
Origin-school signal	(1) Full	(2) LOFO	(3) Full	(4) LOFO
Raw mean Grade 7 grade	+0.310*** (0.021)	+0.109*** (0.022)	+0.023 (0.015)	+0.020 (0.015)
Raw mean Grade 8 grade	+0.223*** (0.032)	+0.233*** (0.034)	+0.601*** (0.021)	+0.572*** (0.022)
Wedge $\hat{\alpha}_j$	−0.069* (0.030)	+0.102*** (0.029)	+0.165*** (0.018)	+0.136*** (0.018)

In the full sample the raw Grade 7 grade predicts +0.310 on the destination's Grade 7 intake (column 1), but only +0.109 once a mover's own feeder-mates leave her peer mean (column 2): the like-for-like advantage was co-movers carrying their shared origin's grades into it. On the Grade 8 achievement margin (columns 3 and 4) the raw Grade 7 grade is within sampling error of zero with or without own feeders (+0.020 under LOFO). Taken at face value, a strong entry grade places a student beside like-graded peers and says nothing about what the cohort she joins goes on to attain. Read alongside its school's wedge, it does: fitted together, the raw Grade 7 mean and the wedge both predict the cohort's Grade 8 attainment, in the ratio the calibration assigns them (Table 46).

The wedge behaves the other way. The operation that collapses the raw Grade 7 coefficient turns the wedge's intake coefficient from −0.069 to +0.102, and barely moves its achievement coefficient, from +0.165 (column 3) to +0.136 (column 4). This does not yet show that the wedge measures learning. Predictor and outcome are both built from Grade 8 grades, and LOFO leaves that shared construction in place; Appendix N.5 breaks it, Appendix N.6 avoids it, and the student-level test of §4.3 of the main text never had it.

The raw Grade 8 grade, output predicting output, survives the same de-leaking almost intact, +0.601 down to +0.572. Its survival shows the operation removes the own-feeder channel and nothing else.

N.4: The negative intake coefficient: a construction artefact

The one negative estimate in Table 39, the wedge's full-sample coefficient on raw Grade 7 intake (−0.069, $t = -2.3$), is arithmetic rather than a preference for weakly graded destinations. By the bridge identity of §3.4 of the main text, $\hat{\alpha}_j$ is to a first order $\overline{y}_j^{G8} - \beta_1 \overline{y}_j^{G7}$, so a high-wedge school has low raw Grade 7 grades by construction, the two correlated at −0.469 in this sample; and the full-sample peer mean keeps a mover's feeder-mates, who carry those low grades into it. Her destination's Grade 7 peer mean is pulled down by her own origin, and that pull is the negative coefficient.

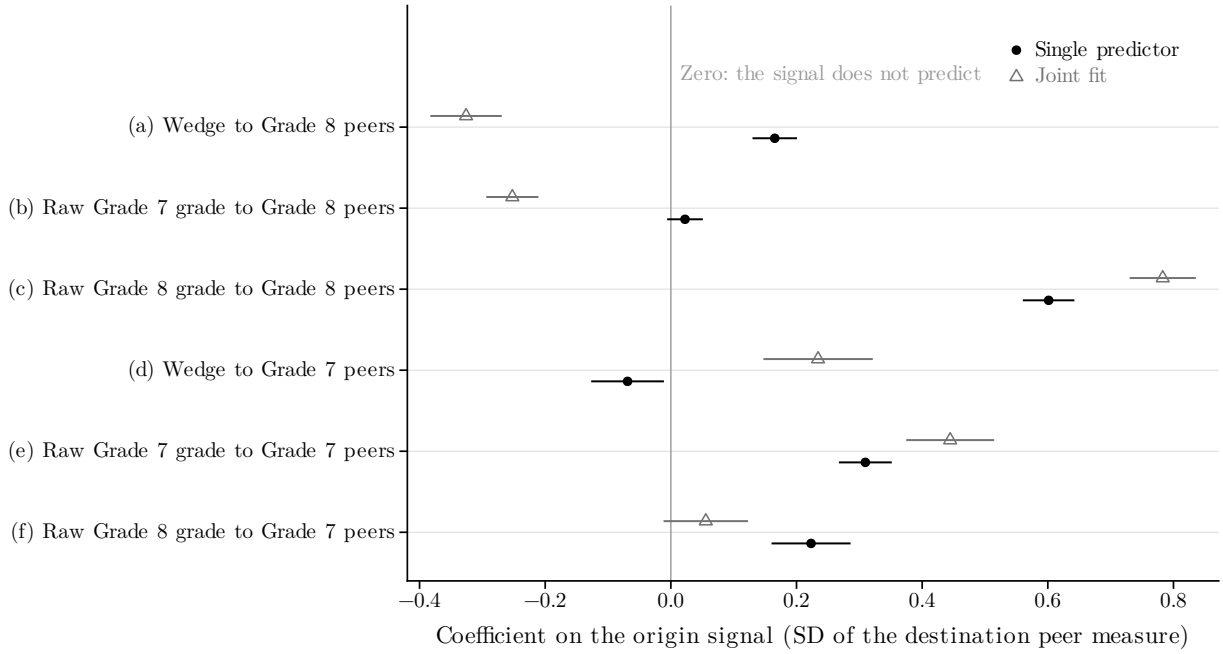

Figure 17: **Each origin signal as a predictor of the destination cohort's Grade 7 and Grade 8 peer means.** *Coefficient and 95% confidence interval (from two-way-clustered standard errors) on each of the three origin-school signals, under single-predictor fits (rows a to f) and a joint specification 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 grade net of a fraction of the Grade 7 grade, conditioning on both raw means reverses its coefficients' signs. Table 45 reports the joint coefficients in full.* ***Sample:*** *221,093 student-subject-cohort observations; 1,371 origin / 883 destination schools; Gauteng; Mathematics; 2018 and 2022 cohorts.*

Table 40: **The wedge's negative intake coefficient is a construction artefact.** *The wedge's full-sample coefficient on the destination's raw Grade 7 intake, the negative estimate in Table 39, split additively into its own-feeder and cross-feeder components. The last column is each component as a share of the full coefficient; shares exceed 100% and turn negative because the two offset. Controls, clustering, intervals and stars as in Table 39.* ***Sample:*** *n = 221,093; Mathematics; Gauteng; 2018 and 2022 cohorts.*

Origin-school signal	Coefficient	(SE)	95% CI	% of total
Full (leave-one-student-out)	−0.069*	(0.030)	[−0.127, −0.011]	100
Own-feeder component	−0.159***	(0.040)	[−0.238, −0.080]	231
Cross-feeder component	+0.091	(0.047)	[−0.002, 0.183]	−131

Table 40 splits the coefficient into the part carried by same-feeder co-movers and the part carried by every other peer. The own-feeder part is −0.159 ($t = -4.0$), 231% of the total and larger than the whole of it; the cross-feeder part, from peers who share none of her origin, is positive (+0.091, $t = +1.9$). Among peers from other schools a higher wedge predicts a higher raw Grade 7 intake, as it predicts higher Grade 8 achievement, and dropping own feeders turns the coefficient positive (+0.102).

N.5: Threats to validity and robustness

Unobserved ability. The regression conditions on the raw Grade 7 grade and on demographics, not on every dimension of Grade 7 ability, so if families also sort on an unobserved ability margin correlated with the wedge the coefficient inherits it. The estimate is permissive against sorting that runs with the observed pattern and conservative only against sorting that runs against it. The sharper form of the concern is structural. The wedge is the school-level part of a student's Grade 8 attainment that her Grade 7 grade does not predict, and the outcome is the realised Grade 8 attainment of the cohort she joins, so predictor and outcome are functions of the same quantity; any process that groups entrants more finely than the Grade 7 grade, whether placement testing, streaming or residential sorting, seats students from high-wedge origins beside higher-attaining peers with no household reading anything. LOFO does not reach this channel, which runs through a student's own attainment rather than her feeder-mates'. Cross-cohort substitution does: replacing her own-cohort wedge with the adjacent cohort's takes her own attainment out of the predictor. The Grade 8 coefficient survives it at +0.108 ($t = +5.5$) and the intake coefficient attenuates to zero (+0.027, $t = +1.06$), as the construction reading of Appendix N.4 predicts. Because the substitution attenuates the behavioural channel along with the mechanical one, the specification that tells them apart severs both at once: cross-cohort estimation on the LOFO outcome leaves the Grade 8 coefficient at +0.095 ($t = +4.8$). The student-level test of §4.3 of the main text does not have this problem. Its outcome is the student's own school-leaving result, recorded five years later in a different record system and standardised nationally, and it compares students within the secondary school they entered, so the outcome contains none of the Grade 8 attainment the wedge is built from. The wedge's coefficient survives there too.

Destination grading. A destination's Grade 8 grades carry that school's own grading, so a student who joins a graded-up cohort appears to join a high-attaining one. Netting the first stage's destination fixed effects out of the outcome raises the coefficient rather than lowering it, to +0.231 ($t = +11.3$ on the same-cohort LOFO outcome). That outcome has lower variance and the destination effect mixes difficulty with instruction, so its level is read with care; its sign runs against the result rather than with it. The specification that severs estimation overlap, own-feeder composition and destination effects together holds at +0.136 ($t = +7.0$).

Residential clustering. Active choice and residential clustering are not separable here: high-wedge primaries may feed particular secondary schools because the two share a neighbourhood. The school-geography bound of Appendix N.12 removes the mechanical nearest-school channel, since the coefficient holds among movers who cross a ward or a municipal boundary, but relocation into a better-served area is observationally equivalent to active choice (Bayer et al. 2007; Rothstein 2006), so the bound sharpens the reading without settling it. None of the three threats bears on the comparison between signals. Raw grades and the wedge identify different sorting margins, with opposite signs on intake, whichever mechanism produces the pattern.

Two checks concern measurement rather than interpretation. Restricting to origins with raw Grade 7 means in [30%, 70%], which removes floor and ceiling compression, retains 95.7% of observations and leaves the wedge at +0.166 on Grade 8 ($t = +9.2$) and −0.057 on intake ($t = -2.0$), a 17.7% attenuation of the intake coefficient consistent with mild ceiling compression rather than a reversal. Only one specification reverses the calibrated signs, and it is a bad control rather than a horse race: entering all three signals jointly conditions on $\overline{y}_j^{G8}$, which contains $\hat{\alpha}_j$ by the bridge identity, and partialling a measure against its own components turns the calibrated coefficient negative on Grade 8 and positive on intake (Figure 17; tabulated in Appendix N.9). Estimated province by province, the wedge's Grade 8 LOFO coefficient is largest in Gauteng; across the six provinces it is about half Gauteng's (Table 42).

N.6: The destination's published matric, an outcome families can see

The headline test pairs two quantities no household reads directly. One academic signal is public at this margin: each destination secondary school has a National Senior Certificate (NSC) record, historically published by school and reported in the press every January, the one observable academic attribute of the schools a family chooses between. The same test run on that public attribute pairs the wedge with a signal households could have read.

The wedge predicts arrival at a destination with a stronger published matric record. Pooling the two cohorts (209,666 movers, 756 destination secondary schools), a student from an origin whose wedge is one standard deviation higher reaches a secondary school whose published NSC Mathematics mean is +0.144 standard deviations higher ($t = 7.1$; Table 41). Raw Grade 7 selectivity, the proxy a study without a calibrated measure would reach for, predicts the same outcome at +0.049; the wedge's coefficient is about 3 times as large, and the two 95 per cent intervals, [0.104, 0.184] and [0.019, 0.078], do not overlap. Raw Grade 7 is small and positive here, where on the peer-Grade-8 margin of Appendix N.3 it was zero.

Table 41: **The wedge predicts the destination's published matric result; raw Grade 7 selectivity predicts it about 3 times more weakly.** *Coefficient (two-way clustered SE) on each standardised origin-school signal from separate single-predictor fits; the outcome is the destination secondary school's published NSC Mathematics mean, standardised within cohort. The matric mean predates the choosing student's arrival, so it carries no own-feeder composition. Controls and province-by-cohort fixed effects as in Table 39. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.* ***Sample:*** *$n = 209{,}666$ movers; 756 destination secondary schools; Mathematics; Gauteng; pooled 2018 and 2022 cohorts.*

Origin-school signal	Coefficient (SE)	95% CI
Wedge $\hat{\alpha}_j$	+0.144*** (0.020)	[0.104, 0.184]
Raw mean Grade 7 grade	+0.049** (0.015)	[0.019, 0.078]
Raw mean Grade 8 grade	+0.224*** (0.027)	[0.171, 0.277]

This corroboration runs through a channel the construction artefact cannot reach. A published matric mean is earned by cohorts who left years before the choosing student arrived, so it cannot contain her feeder-mates, and the matric result reproduces the wedge's predictive content with none of the leave-own-feeder-out machinery. It holds in each cohort (2018 +0.143, 2022 +0.163), for the school's overall NSC mean across subjects (+0.183, the closest match to the reputation a family hears), and forward to the destination's own matric output four years after Grade 8 entry (+0.134, $t = 5.2$); Appendix N.13 reports every estimate. A matric mean reflects whom a school admits as much as what it teaches, so the result does not show that the chosen destinations produce more. A quarter of the matric score is also the destination's own grading: the NSC blends school-based assessment, set and graded by the Grade 12 school, at 25% with an external examination at 75%. That quarter carries the same kind of difficulty this paper measures. If primaries with harder assessments send their leavers to secondary schools whose school-based assessments are also harder, as the Grade 8 evidence of Appendix N.5 suggests they do, the harder assessments lower the destinations' matric means and pull the coefficient down; if they send them to secondary schools with easier assessments, it pushes the coefficient up. Nothing here bounds that quarter. The revealed-preference reading needs only that families move towards schools that are publicly stronger, whatever makes them so.

N.7: Scope restriction and sample construction

The tests of this appendix run in Gauteng, 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; §4.8 of the main text reports the main tests across the six provinces, with Gauteng's beside. English first additional language is excluded because within-year reliability falls below the bar set in §4.2 of the main text. Table 43 traces the sample from all Grade 7 to Grade 8 mathematics movers to the Gauteng 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 Gauteng, which leaves 221,093 student-subject-cohort observations.

Table 42 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; across the six provinces the coefficient is about half Gauteng's.

Table 42: 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 Appendix N.4.

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 row sits a few thousandths below Gauteng's own-sample estimate in Table 39; the final row is the six provinces, the figure §4.8 of the main text gives.

Table 43: Sample attrition: mathematics, Gauteng. *Sample:* each step's own, from all switchers (all provinces and subjects) to the full sample of Table 39: Mathematics; Gauteng; 2018 and 2022 cohorts.

Step	n
All switchers (non-missing identifiers, Grade 7 grade; 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	221,093

Table 44: Pairwise correlations among the three origin-school signals (school-level, mathematics, Gauteng). *Sample:* origin schools, at school level; Mathematics; Gauteng.

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

The pattern of predictor correlations matches the school-level correlations reported in Table 44: $\hat{\alpha}_j$ 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.

N.8: Caveats on movers and geographic resolution

Two named caveats supplement the three caveats in §4.8 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 §4.8 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 sorting strength of the simulations (Table 21) is silent on the cross-level distinction (within- vs across-province sorting); the sorting test here covers 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).

N.9: Horse-race specification

Figure 17 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 Appendix N.3 beside the intake result of Appendix N.4. 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 Appendix N.4, while the raw Grade 7 grade sits at +0.310 (row e), the grade-on-grade matching the wedge nets out, and the raw Grade 8 grade 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 Appendix N.5, 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, $\overline{y}_j^{G8}$ contains $\hat{\alpha}_j + \beta_1 \overline{y}_j^{G7}$ up to estimation noise, so conditioning on it mechanically reverses the calibrated partial coefficients: with $\hat{\alpha}_j$ negatively correlated with $\overline{y}_j^{G7}$ and positively correlated with $\overline{y}_j^{G8}$ (Table 44), both calibrated coefficients in Table 45 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 estimates, and the construction-robust decomposition is in Appendix N.10.

Table 45: Horse-race specification: all three origin-school signals entered simultaneously. Same controls and clustering as Table 39. * $p < 0.001$. ***Sample:* the full sample of Table 39: Mathematics; Gauteng; 2018 and 2022 cohorts.

Origin-school signal $\mathcal{S}_j$	Dest. peer G7 grade (z)	Dest. peer G8 grade (z)
$\overline{y}_j^{G7}$ (mean Grade 7 grade, z)	$+0.445^{***}$ (0.036)	-0.252^{***} (0.021)
$\overline{y}_j^{G8}$ (mean Grade 8 grade, z)	$+0.056$ (0.034)	$+0.783^{***}$ (0.027)
$\hat{\alpha}_j$ (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 46). 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 grade 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 46: **The raw Grade 7 mean and the wedge enter in the ratio the calibration assigns**

	Leave-own-feeder-out			Own-pupil-out
	2018	2022	Pooled	Pooled
	(1)	(2)	(3)	(4)
Raw Grade 7 mean	0.216***	0.171***	0.181***	0.217***
	(0.033)	(0.025)	(0.022)	(0.021)
Wedge $\hat{\alpha}_j$	0.288***	0.277***	0.269***	0.324***
	(0.037)	(0.034)	(0.027)	(0.025)
Own Grade 7 grade	0.068***	0.066***	0.069***	0.068***
	(0.008)	(0.010)	(0.009)	(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 grade, 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, 2018 and 2022 cohorts: the leave-own-feeder-out sample of Table 48, 220,973 student-subject-cohort observations (94,972 in 2018, 126,001 in 2022); in column (4), the full sample of Table 39, 221,093.*

N.10: Same-feeder exposure and the leave-own-feeder-out re-estimate

The destination peer means in the full-sample columns of Table 39 are leave-one-out at the *student* level: a mover's own grade 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 grades, same-feeder co-movers carry origin $\hat{\alpha}_j$ into the outcome mechanically. Table 47 reports the exposure: the share s_i of each student's leave-one-out peer pool drawn from her own feeder.

Table 47: Same-feeder share s_i of the leave-one-out destination peer pool. Sample: $n = 221,093$ student-subject-cohort observations (the full sample of Table 39); mathematics; Gauteng.

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 48 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 × cohort, which carries a destination's assessment difficulty together with its teaching (Appendix N.5). The cross-cohort × LOFO row composes this exclusion with Appendix N.11's design, severing the estimation-overlap and composition channels simultaneously. The paired-cluster bootstrap for the LOFO Grade 8 coefficient ($B = 999$; origin clusters resampled with replacement, $\hat{\alpha}$ redrawn per replicate) gives the 95% interval reported below.

Table 48: Leave-own-feeder-out re-estimates and the exact same-/cross-feeder decomposition. Same controls and clustering as Table 39. Sample: $n = 220,973$ (destinations with ≥ 5 cross-feeder mover peers on both margins; 99.9% of the full sample); mathematics; Gauteng; two-way clustered (origin, destination). *** $p < 0.001$.

Coefficient	Estimate (SE)	t	n
$\hat{\alpha}_j \to$ dest. G8, leave-one-student-out (headline)	$+0.165^{***}$ (0.018)	+9.2	221,093
$\hat{\alpha}_j \to$ dest. G8, LOO on LOFO subsample	$+0.165^{***}$ (0.018)	+9.1	220,973
$\hat{\alpha}_j \to$ dest. G8, **leave-own-feeder-out**	$+0.136^{***}$ (0.018)	+7.4	220,973
$\hat{\alpha}_j \to$ dest. G8, γ_k-netted LOO	$+0.273^{***}$ (0.020)	+13.8	221,093
$\hat{\alpha}_j \to$ dest. G8, γ_k-netted LOFO	$+0.231^{***}$ (0.020)	+11.3	220,973
$\hat{\alpha}_j \to$ dest. G7, leave-one-student-out (headline)	-0.069 (0.030)	-2.3	221,093
$\hat{\alpha}_j \to$ dest. G7, leave-own-feeder-out	$+0.102^{***}$ (0.029)	+3.6	220,973
$\overline{y}_j^{G7} \to$ dest. G8, LOFO (placebo)	+0.020 (0.015)	+1.4	220,973
$\overline{y}_j^{G7} \to$ 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
Cross-cohort × LOFO, $\hat{\alpha}_j \to$ 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-grade benchmark points the same way without assuming the result. Because flows are feeder-clustered, an allocation on raw Grade 7 grades alone, with no household sorting, would still produce a positive same-feeder component, while its cross-feeder prediction is zero only insofar as school-level raw grades carry no cross-school achievement signal. Both halves hold in the data: cross-feeder matching on raw grades 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 48.

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.

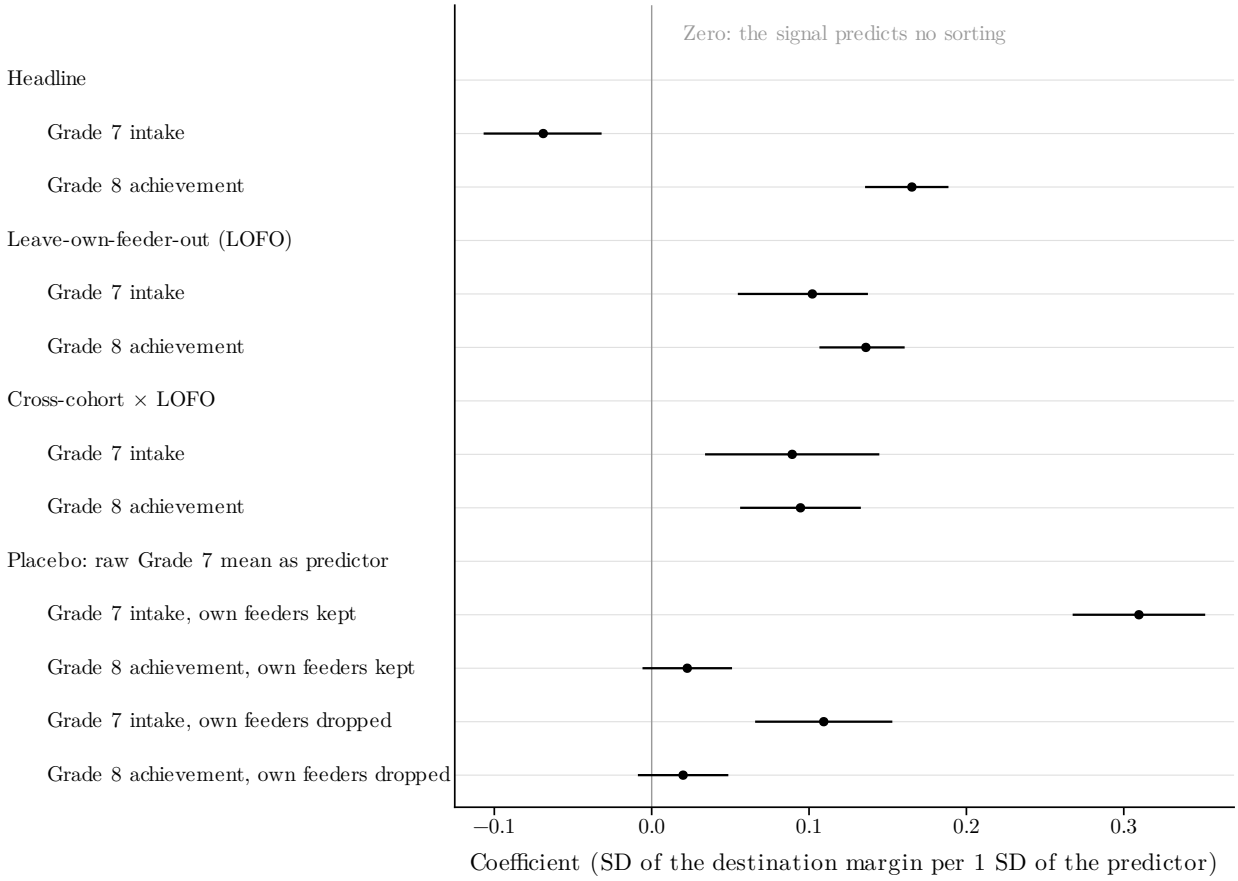

Figure 18: **Every specification keeps the wedge's achievement coefficient; the raw-grade placebo on achievement is near zero.** *Coefficient on the origin-school signal across four specifications (headline, leave-own-feeder-out (LOFO), cross-cohort × LOFO, and a placebo using the raw Grade 7 mean as the predictor), on both destination margins. The placebo is near zero on the Grade 8 achievement margin whether or not own feeders are dropped; its large coefficient on the Grade 7 intake margin is the like-for-like matching decomposed in Appendix N.4. LOFO intervals are bootstrapped. The one point left of zero is the headline wedge coefficient on Grade 7 intake, the construction artefact of Appendix N.4; it turns positive once own feeders are dropped.* ***Sample:*** *LOFO subsample n = 220,973 (99.9% of the full sample); Gauteng; Mathematics; 2018 and 2022 cohorts.*

N.11: Cross-cohort alpha robustness

Gauteng's pooled cross-cohort estimate 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-grade outcome. The pooled estimate of +0.108 ($t = 5.5$) is a precision-weighted combination; because the 2018-choices/2022-alpha estimate 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 49: 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 full sample whose origin school has a wedge in the other cohort; Mathematics; Gauteng; 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

N.12: 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 subsection 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 50 reproduces Gauteng's estimate in Table 39 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 50: 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 full sample); mathematics; Gauteng, 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)

(continued)

Specification	n	Dest. G7 peer (z)	Dest. G8 peer (z)
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)
$\widehat{\alpha}_j\times$ distance (per +100 km)	215,209	$+0.650^{***}$ (0.084)	-0.203^{***} (0.056)
Marginal $\widehat{\alpha}_j$ at median distance (1.7 km)	215,209	-0.103^{***} (0.030)	$+0.161^{***}$ (0.019)
Marginal $\widehat{\alpha}_j$ 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)
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 50). It weakens with distance but stays positive across the observed range: the $\widehat{\alpha}_j\times$ 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 estimate 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 enrols locally produces exactly the long-displacement, boundary-crossing pattern just described, and that pattern is *strongest* in precisely the cross-municipality sample 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 §4.8 of the main text stands: the central revealed-preference claim is the equilibrium one, that raw grades 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 $\widehat{\alpha}_j$.

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 $\widehat{\alpha}_j$ (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 coefficient (§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.

N.13: Published matric: an own-feeder-immune corroboration

Appendix N.6 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 regression is that of §4.8 of the main text, 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 estimates before any matric estimate 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 51 collects the single-predictor coefficients across cohorts and outcome variants. The calibrated standing predicts the destination's published matric in every cohort and outcome variant, 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 N.10 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 N.9) leaves it positive here, at +0.248 ($t = 6.0$).

Table 51: 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 grade 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; 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 O: Value added

With the calibrated score as the intake, a secondary school's effect on the school-leaving result is how far its students finish above students who entered with the same score. Such effects could still record who enrols, how classmates shape each other, or how generously a school grades the quarter of the NSC it sets itself. The calibrated intake passes the standard placebo test that the raw intake fails: on the Grade 7 Term 1 grade, earned before entry, the effects disperse at 0.364 of their own dispersion with the calibrated intake, against 0.696 with the raw (Appendix O.4), and moderation bounds how far a school's own grading can lift its effect (§4.9 of the main text). The effects remain descriptive value added: a composite of selection,

instruction and peer effects. Appendix O.1 lists every check with its threshold and result, the subsections after it give the design and the checks in full, and Appendix O.9 shows why the intake has to be a grade the secondary school did not set.

O.1: Checks on the exit-school effects

This subsection lists the checks behind §4.9 of the main text.

Table 52: **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 correlation instead
Signal	signal-to-noise ratio	2.450	2.403	≥ 2.0	passes	not reported
Specification	lowest agreement across choices that leave the intake unchanged	0.972	0.973	≥ 0.80	passes	outcome definitions varied
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 grade	dispersion ratio, raw Grade 7 Term 1 grade	1.061	1.282	≤ 1.00	fails	not reported
Placebo, calibrated grade	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

(continued)

Check	Statistic	Mathematics	Mathematical Literacy	Threshold	Result	Liu et al. (2025)
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
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.

O.2: The intake problem

The choice of intake matters: a placebo that no secondary school could have caused, its entrants' Grade 7 Term 1 Mathematics grade read on the intake's own scale, disperses across secondary schools at 0.696 of the Mathematics effects' own dispersion with the raw grade as the intake and at 0.364 with the calibrated score, and 40.4% of secondary schools change quintile of the Mathematics effects between the two intakes. To move from descriptive to causal value added takes two things: comparable measures of learning at entry and exit, and assignment to secondary schools that is as good as random, such as an admission lottery (Deming 2014; Angrist et al. 2017). The calibrated score supplies the first at entry, where no common test exists; at exit, three quarters of the school-leaving score is an external examination and the other quarter is graded by the student's own Grade 12 school (Appendix O.6). This setting lacks the second.

O.3: Design

For student i entering secondary school k,

$$Y_i = \theta_k + \delta A_i + X_i'\varphi + \varepsilon_i,$$

fitted separately in each of six provinces. Y_i is the student's 2025 National Senior Certificate (NSC) score in Mathematics or in Mathematical Literacy, standardised nationally, so θ_k is in national standard deviations of that score. A_i is the calibrated score of §3 of the main text, the student's Grade 7 Mathematics grade net of her primary school's wedge, and X_i holds the controls of §4.3 of the main text. θ_k, the exit-school effect, is how far school k's entrants finish above or below what their intake and controls predict, relative to the province's average secondary school. Mathematics comes first throughout, with Mathematical Literacy, estimated separately, in parentheses.

The regression is the one in §4.3 of the main text in which the calibrated score is the only Grade 7 regressor, on the same 394,353 students; there I compare its fit with the raw grade's, and here I fit it province by province for its fixed effects. It carries the limits set out at the end of that section; the one that bears more heavily on a school's effect than on a slope, the blended outcome, is taken up in Appendix O.6. Every statistic below is computed over the 1,135 (1,486) secondary schools with at least twenty sitters of the subject.

The design mirrors that of Liu et al. (2025), who estimate a sending school's effect on a school-set outcome, Grade 9 grade-point average, with lagged state test scores as the intake and a receiving-school fixed effect to absorb the destination's grading. Here three quarters of the outcome is an external examination and no external intake exists, so the calibration does the work that the state test does there.

The standard check on an intake is a placebo test: refit the estimator with a grade earned before entry as the outcome. No secondary school can have changed its entrants' Grade 7 grades, so school "effects" on such a grade that disperse as widely as θ_k mean that the effects record who enrols. The placebo is the Grade 7 Term 1 Mathematics grade, three terms older than the Term 4 grade the intake is built from, and the statistic is the ratio of the two signal standard deviations, each in units of its own outcome. Like the intake, the placebo can be read raw or calibrated.

O.4: The placebo test

The raw intake fails the placebo test and the calibrated intake passes it. With the raw Grade 7 grade as the intake, the raw Term 1 grade disperses across secondary schools at 0.696 (0.704) of the effects' own dispersion, above one half. With the calibrated score as the intake, the calibrated Term 1 score disperses at 0.364 (0.434), under half, and its disattenuated rank correlation with θ_k is −0.161 (−0.205), so the secondary schools whose

entrants finish ahead are not the ones whose entrants started ahead. The two fits share students, schools and estimator, and differ only in whether the Grade 7 grades still carry the primary school's wedge.

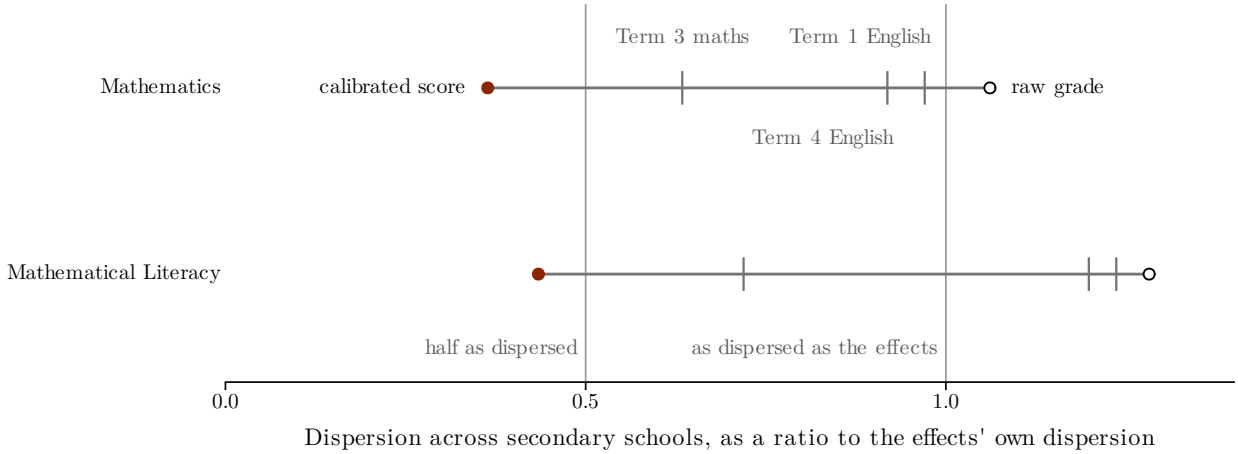

Figure 19: **Against a calibrated intake, the pre-entry grade passes the placebo test calibrated and fails it raw.**

Notes: Each marker is the dispersion across secondary schools of the school effects on a Grade 7 grade that no secondary school could have caused, as a ratio to the dispersion of the exit-school effects θ_k, for Mathematics and for Mathematical Literacy estimated separately. Dispersion is the signal standard deviation of a set of school effects, the square root of the covariance between the effects estimated on two random halves of each school's students, with each set divided by the standard deviation of its own outcome, so the ratio has no unit. In every fit drawn, the intake is the calibrated Grade 7 Mathematics score alone; the fits with the raw grade as the intake are reported in the text and are not drawn. The open circle is the raw Grade 7 Term 1 Mathematics grade as the placebo and the filled circle the calibrated score from the same grade; refitted on exactly the calibrated score's students and schools, the raw ratio is unchanged at three decimals. The grey ticks are three held-out calibrated placebos, none of them in the intake: the calibrated scores from the Grade 7 Term 3 Mathematics grade and from the Term 4 and Term 1 English First Additional Language grades. The two grey vertical lines, drawn behind the data, mark a placebo as dispersed as the effects, at one, and half as dispersed, at one half. The outcome is the NSC score, the final blended score: a quarter school-based assessment and three quarters external examination. The missing-value glyph for this document is an em dash. None arises here. Sample: the cohort in Grade 7 in 2020, Grade 8 in 2021 and the NSC in 2025, in six provinces (Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga and North West), 394,353 linked students at secondary schools with at least twenty linked students holding a calibrated Grade 7 intake; each Mathematics ratio is computed over the 1,135 secondary schools with at least twenty Mathematics sitters and each Mathematical Literacy ratio over the 1,486 with at least twenty Mathematical Literacy sitters.

Figure 19 holds the intake at the calibrated score and varies only the placebo's scale. Read raw, the same Term 1 grade disperses at 1.061 (1.282), more widely than the effects themselves, because it still carries the primary school's wedge, which the calibrated intake has removed, and the wedge reappears as a difference between the secondary schools those primaries feed. Against the raw intake, the calibrated placebo fails for the mirror-image reason, at 1.122 (1.263). A placebo on a school-set grade tests an intake only when the two share a scale.

O.5: Why the raw intake fails

The gap between the raw-intake and calibrated-intake effects tracks the mean wedge of each school's entrants, which accounts for 83.4% (85.2%) of its variance, although the gap rises with the wedge at only 0.62 (0.54) of the rate the intake coefficient implies. The two sets of effects agree at a disattenuated rank correlation of 0.877 (0.867), yet 40.4% (44.2%) of secondary schools change quintile between them (Mathematics in Figure 13 of the main text). The raw-intake placebo understates the error, because a raw placebo carries the same wedge as a raw intake and the two partly cancel, whereas the examination component carries none.

The intake nearest to hand does worse. A secondary school's own Grade 8 grade is the latest measure of its entrants, but the school sets it, so a school whose Grade 8 assessments are easy, or that teaches its entrants well in their first months, makes them look stronger and is charged for it through the intake coefficient.

With that grade as the intake, the effects move with the school's destination effect $\hat{\gamma}_k$ at −0.498 (−0.405) national standard deviations per unit of the destination effect, one within-province standard deviation of Grade 8 grades, against −0.462 (−0.474) implied by the intake coefficient; the destination effect accounts for 81.8% (76.1%) of the change, and 54.0% (54.4%) of secondary schools change quintile. Appendix O.9 shows the same problem in a national school-based-assessment extract, where a single school fixed effect, net of an intake grade the school set, recovers the sum of how hard its intake assessments were and how far its students finish above that intake. An intake set by other schools and placed on a common scale keeps the school's own Grade 8 grading out of θ_k, and leaves its early teaching in.

O.6: Reliability, and what it cannot separate

The effects are precisely estimated: split-half reliability 0.855 (0.841), signal-to-noise ratio 2.45 (2.40), and a corrected coefficient of 1.030 (1.051) when effects estimated on half of each school's students predict the other half's outcomes.

The extract of Appendix O.9 gives a separate estimate of how hard each secondary school's own assessments are; comparing it with θ_k school by school would show how much of θ_k is grading, and I have not yet made that comparison.

The same extract, in a two-way fit, separates how hard each secondary school's intake assessments were from how far its students then finish above their intake, and the second component has a split-half reliability of only 0.377. That figure does not conflict with θ_k's. The two-way fit can separate the components only through students who change school between the intake grade and the examination, so few students inform each school's estimate, whereas θ_k is a single fixed effect informed by every entrant, which the design can afford because the secondary school did not set the intake.

These checks differ from the evidence for the closest design. Liu et al. (2025) draw on several cohorts, so they report the year-to-year correlation of their school measure and its prediction of later outcomes, grade repetition and college enrolment among them, which one linked cohort cannot supply. Their report carries no placebo on a pre-entry outcome, no split-half reliability, no bound on sorting on attainment the intake does not capture and no test among students who did not choose their school. This design carries all four because, without an external intake, it has to show that the calibrated one removes the sorting a raw grade lets through.

O.7: What the calibrated intake leaves behind

Grades the intake has not seen show what it misses. With the calibrated Term 4 Mathematics score alone as the intake, other calibrated Grade 7 scores still disperse across secondary schools: Term 3 Mathematics at 0.634 (0.719) of the effects' dispersion, and English First Additional Language at 0.919 (1.198) in Term 4 and 0.971 (1.237) in Term 1. What a mathematics intake leaves behind runs through language grades. Adding the calibrated Term 3 and Term 4 scores in both subjects, with splines, brings the Term 1 English and Mathematics placebos to 0.440 (0.530) and 0.431 (0.544), under one half for Mathematics and just over it for Mathematical Literacy, and moves θ_k by 0.143 (0.281) of its signal standard deviation. The splines, the last of those steps, move it by 0.075 (0.064), adding the two Term 1 scores moves it by 0.014 (0.034) more, and successive steps agree above 0.99.

The residual is small against the effects' signal. The calibrated Term 1 Mathematics placebo accounts for 1.8% (2.8%) of the effects' signal variance, 18.6% (29.6%) in a one-for-one worst case. Adjusting θ_k for the measured sorting leaves 95.1% (95.1%) of secondary schools in the same band, above, indistinguishable from or below the province average, and the proportional-selection bound of Oster (2019), which scales unmeasured sorting with measured sorting, leaves 78.0% (77.7%). What remains follows geography more than choice: among students who did not choose their secondary school, the calibrated placebo disperses more widely, at 0.675 (0.688) against 0.431 (0.544) for all students with the same intake. A school-level residual from the cohort that wrote the examination in 2023 agrees with the effects at 0.518 (0.536); it carries no student controls and falls short of 0.60, the agreement a student-level cross-cohort test would need. The shortfall comes mostly from change between the cohorts, not from the cruder object. Built the

same way for the cohort that wrote in 2025, the residual correlates 0.908 (0.905) with that cohort's effects but only 0.619 (0.608) with the 2023 residual, both corrected for noise.[4]

O.8: What would make the effects value added

A forecast test on a second linked cohort (Chetty et al. 2014) waits for the Grade 7 2021 cohort, which writes the examination in November 2026; its panel is built, with 5,169 secondary schools holding twenty students in both cohorts, and the Grade 7 2018 cohort's 2023 examination is not linked to students. An external Grade 7 assessment would allow the comparison Liu et al. (2025) make against test-score value-added, and South Africa has had none since 2014. Assignment variation, such as Gauteng's feeder-zone admission rules, would allow a causal reading; no source confirms that the records exist. The comparison with each secondary school's own grading can run now.

O.9: What a single school fixed effect on the examination recovers

Figure 20 reports 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 comparisons, 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 comparison is the NSC score, the final blended score of school-based assessment and an external examination. The exit-school effect is not offered as an ordering of schools. Table 53 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 Gauteng's Grade 7 2018 Mathematics cohort. The intakes are at Grades 8, 9 and 10, the subjects are 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.

[4]Both are covariances over the product of the two vectors' split-half signal standard deviations, on the secondary schools present in both cohorts. On the same estimator and schools the 2023 residual agrees with the effects at 0.558 (0.585), close to the product of the two.

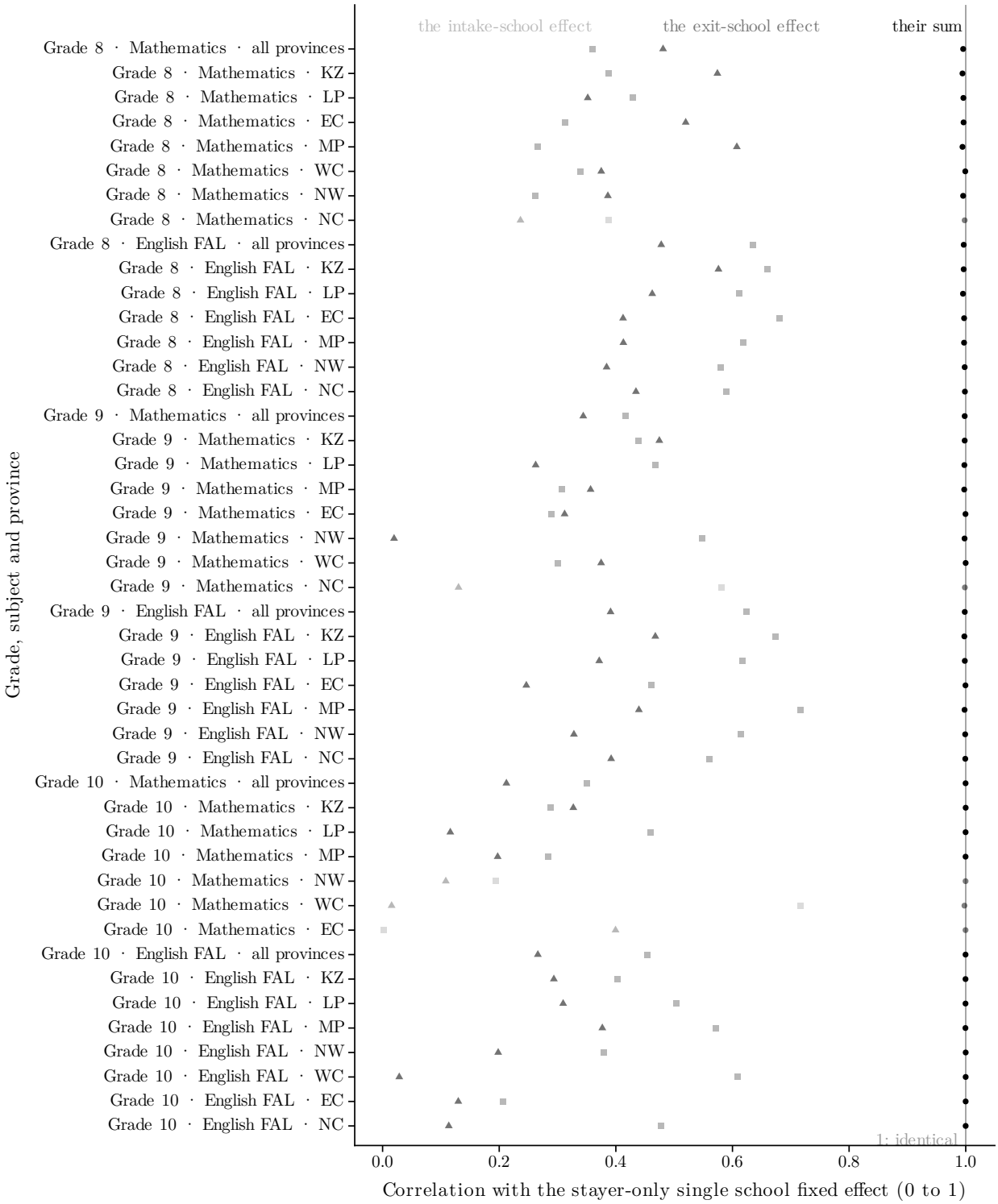

Figure 20: **A single school fixed effect recovers the sum, not either component.**

Note. Each row is one comparison: 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 grade; the intake-school effect is the fixed effect on the school that set the intake assessment, the sign of the paper's wedge. The National Senior Certificate score is the final blended score, 25% school-based assessment plus 75% external. Neither effect is a causal value-added estimate. Across the 45 comparisons 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 row 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 1,320 schools, against 0.939 for Gauteng's wedge over 1,414 schools, as the table below reports.** No significance star appears in this figure, and none applies; 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; one further row of the source compares too few schools to carry a correlation at all and is not plotted. Source: analyses/sba_nsc_link/outputs/tables/tab5_design_b.csv, via outputs/tables/mathematics/tab_nsc_single_fe_recovers_sum.csv. **Sample:** students with a valid Grade 8, 9 or 10 Term 4 2021 school-set grade at a school of at least ten such students and a same-subject score at their own on-time National Senior Certificate, exit province equal to intake province; the 45 rows compare 30 to 3,716 schools each.

Table 53: **The exit-school effect reproduces far less well than the wedge.**

	Reproducibility			Dispersion		
	Schools	Split-half	Reliability	SD	SE	Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. This extract, the exit-school effect at Grade 8						
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]
Panel B. The wedge in Gauteng						
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 grade for the same pupils, in within-province standard deviations of that score. The National Senior Certificate score is the final blended score, 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 grade. 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 Gauteng's 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 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, 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 lists where each number comes from. **Sample:** Panel A, pupils with a valid Grade 8 Term 4 2021 school-set grade 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 Mathematics panel of 1,414 schools in the 2018 cohort.

Appendix P: Replication

P.1: 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.

P.2: 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 grades, 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.

References

Abdulkadiroğlu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters. 2020. "Do Parents Value School Effectiveness?" *American Economic Review* 110 (5): 1502–39. https://doi.org/10.1257/aer.20172040.

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.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *American Economic Review* 107 (6): 1535–63. https://doi.org/10.1257/aer.20140774.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics* 3 (3):

29–54. https://doi.org/10.1257/app.3.3.29.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics* 132 (2): 871–919. https://doi.org/10.1093/qje/qjx001.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115 (4): 588–638. https://doi.org/10.1086/522381.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *The Quarterly Journal of Economics* 114 (2): 577–99. https://doi.org/10.1162/003355399556070.

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.

Deming, David J. 2014. "Using School Choice Lotteries to Test Measures of School Effectiveness." *American Economic Review* 104 (5): 406–11. https://doi.org/10.1257/aer.104.5.406.

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.

Hastings, Justine S., and Jeffrey M. Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." *The Quarterly Journal of Economics* 123 (4): 1373–414. https://doi.org/10.1162/qjec.2008.123.4.1373.

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.

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.

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.

Mizala, Alejandra, and Miguel Urquiola. 2013. "School Markets: The Impact of Information Approximating Schools' Effectiveness." *Journal of Development Economics* 103: 313–35. https://doi.org/10.1016/j.jdeveco.2013.03.003.

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.

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.

Rothstein, Jesse. 2006. "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition Among Jurisdictions." *American Economic Review* 96 (4): 1333–50. https://doi.org/10.1257/aer.96.4.1333.

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." Working paper. February. https://desislavatartova.github.io/pdf/Network_TVA_Desislava_Tartova.pdf.

Woolman, Stu, and Brahm Fleisch. 2006. "South Africa's Unintended Experiment in School Choice: How the National Education Policy Act, the South African Schools Act and the Employment of Educators Act Create the Enabling Conditions for Quasi-Markets in Schools." *Education and the Law* 18 (1): 31–75. https://doi.org/10.1080/09539960600787317.