

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

Job Market Paper

Peter Courtney*

This version: 22 October 2026

Latest version: courtney.ac/comparable-marks

Abstract

Assessment has been found to raise learning system-wide where its results can be compared across schools, yet at most a quarter of countries test every primary student comparably. I recover a common scale of learning outcomes—without a new test—from marks that are incomparable because assessments are more difficult at some schools than at others. Each secondary school sets the same assessments for students from many primary schools within months of arrival, allowing comparison of those primary schools’ marks. Primary schools feed many secondary schools, enabling a two-way fixed-effects estimator to chain this network of comparisons across 1.2 million South African students into a common scale. The gap between rich and poor schools halved on raw marks between 2018 and 2022 but did not narrow on the common scale. Compared with raw marks, about three-quarters of primary schools move to a different quintile of learning. Five years after primary school, the common scale predicts the national school-leaving examination better than raw marks. Easier assessments do not buy a higher-achieving secondary school: at the same raw mark, students whose primary school set easier assessments join lower-achieving classmates. The common scale supplies comparable intake marks, a necessary condition for any value-added measure.

Keywords: school quality; grading stringency; comparability; school-based assessment; two-way fixed effects; South Africa. *JEL:* I21, I24, I28.

1 Introduction

The World Bank’s flagship education review recommends that countries start tackling the learning crisis with “more and better measurement of learning” (World Bank 2018, 17).¹ The evidence suggests the binding word is *better*: Bergbauer et al. (2024) find that introducing comparable national assessment is associated with learning gains of roughly a quarter of a standard deviation of student achievement on international tests, about a year of schooling, and up to twice that in the lowest-performing systems, while “internal reporting that simply informs or monitors progress without standardized comparability has little discernible association with learning outcomes” (2024, 352).² Decisions about learning need measures of learning that mean the same thing at different schools. Comparable results allow parents, teachers, and schools to see where a child

*Stellenbosch University, Tinbergen Institute, and Vrije Universiteit Amsterdam. Email: p.sr.courtney@gmail.com. I thank Debra Shepherd, Servaas van der Berg, Menno Pradhan, and Nic Spaull for their guidance and comments. All errors are my own. The online appendix is at <https://courtney.ac/comparable-marks/>.

¹The review states that policymakers need comparable assessment to “understand whether students are mastering the national curriculum, in which areas students are stronger or weaker, whether certain population groups are lagging behind and by how much, and which factors are associated with better student achievement”.

²Their measure is standardised testing with external comparison (SCOMP), which requires more than comparability alone: the assessment instrument must be set outside the school, and its results must be compared across schools or used for student placement or certification. However, each of the uses highlighted by the World Bank in the previous footnote depends on being able to compare schools, groups, and subjects on a common scale, which requires only that the measures be comparable.

or a school stands against others. They also enable district and provincial administrators to target support, governments to evaluate how a policy or a shock changed learning, and researchers to use learning as an outcome measure.

Yet comparable measurement is rare. At most a quarter of countries test every student on a common instrument at the end of primary school (Gustafsson 2019). Across 63 education systems in Africa and South Asia, a third have no primary leaving examination at all, and several have dropped theirs as secondary places expanded (Rossiter and Konate 2022).³ In most countries, South Africa among them, schools set and grade their own assessments, so the same mark represents different amounts of learning at different schools. The World Bank review concludes that there is “no effective way to aggregate the results of classroom-level formative assessment by teachers into ... reliable system-level information” (World Bank 2018, 18). This paper does exactly that for comparisons between schools in the same year.

Without fielding a new test, I place South African primary schools’ marks on a common scale of learning outcomes by following 1.2 million students across two cohorts as they move from primary to secondary school. Each student has two marks about six months apart: her last mark in primary school and her mid-year mark in her first year of secondary school. I estimate a two-way fixed-effects model that regresses the second mark on the first, with fixed effects for the primary school she left and the secondary school she entered. The secondary school’s effect absorbs everything its new students share there, from how hard its assessments are to what it teaches before June. Because students’ paths between schools vary, both effects can be estimated.

The primary school’s effect is its difficulty wedge, measured in within-province standard deviations (SD) of students’ marks: positive where a school’s assessments are harder than average and negative where they are easier. An assessment is easier if a given mark takes less learning to earn on it, as set and as marked (the data do not separate the two). Reading the primary school’s effect this way assumes that, over those six months, students from primary schools with harder assessments learn as much as those from primary schools with easier ones, and that students do not sort into secondary schools by how fast they are improving. Section 3.3 tests both: wedges estimated from the last term of the first secondary year agree closely with those from its first term (0.902), as they would not if students from primary schools with harder assessments had kept learning faster, and students headed for the secondary schools with the highest and the lowest fixed effects were improving at the same rate in their last primary year (within a hundredth of an SD). Combining each student’s mark with her primary school’s wedge gives her calibrated score, her place on the common scale.⁴ Averaging over a school’s students gives its calibrated level of learning.

The intuition behind the estimator is a comparison between students with the same mark. Two students arrive in the same secondary school having received the same mark from two separate primary schools. Soon after arriving, they sit an identical set of assessments and receive different results. On average, the student who scores higher faced harder assessments in primary school. The comparison also runs the other way: comparing two students who leave the same primary school with the same mark, earned on the same assessments, and enter different secondary schools shows which of those secondary schools gives the lower mark, whether through harder assessments or through what it taught in the months before. Assessments need only be common within each school: the secondary school for the first comparison and the primary school for the second. Chained across the network of transitions, these comparisons put almost every primary school in each province on the common scale.

The data are administrative: student-level records from the South African School Administration and Management System (SA-SAMS), in which schools record their students’ marks, linked from each student’s primary school to her secondary school.⁵ I use the mathematics marks of the cohorts that left primary school in 2018 and in 2022, in six of South Africa’s nine provinces. Students disperse widely across schools during this

³The exam rationed scarce secondary places, so it is dropped where secondary supply has expanded. The systems dropping it are not the most developed (Rossiter and Konate 2022).

⁴The calibration is statistical moderation (Kolen and Brennan 2014, sec. 10.1.2; Masters 1988), not test equating, which requires assessments built to a common specification (Appendix Q).

⁵Primary school ends in Grade 7 and secondary school begins in Grade 8. The two marks are the Grade 7 Term 4 mark, the last report mark of primary school and the one that decides promotion, and the Grade 8 Term 2 mark, from the term in which every Grade 8 student at a school sits the same grade-wide assessment. The two cohorts were in Grade 7 in 2018 and 2022, and in Grade 8 in 2019 and 2023 (Section 2).

transition, with the average secondary school receiving students from 12 different primary schools, which allows each secondary school’s assessments to compare many primary schools at the same time. Schools record a mark in each of the four terms of a school year that runs from January to December, each set and marked by the student’s own school.

Because very few students move across provincial borders, the administrative records are linked within each province and the common scale is estimated separately by province. Unless otherwise specified, all results are reported for Gauteng; the main results are also estimated for each of the other provinces, where they generally hold and are reported in appendices.

Schools’ wedges differ widely, are measured precisely, and persist. Schools at the 10th and 90th percentiles of difficulty mark the identical underlying learning more than a full SD apart. Split each school’s students at random into two halves, and the two estimates of its wedge correlate at 0.88. Four years apart, with new students, new tests, and nothing linking the years, the wedge still correlates at 0.62: it belongs to the school, not to one cohort of its students. The estimator’s two known weaknesses—effects identified only where movers connect schools and variance inflated where the network is sparse (Andrews et al. 2008; Kline et al. 2020)—also matter less here, because almost every student changes school at once and the network connects nearly every school in each province (98.8% or more).

The calibration changes conclusions about socioeconomic achievement gaps, school rankings, secondary-school sorting, and estimates of secondary-school value added. South Africa ranks its public schools into five poverty quintiles by the poverty of the communities they serve. On raw marks, the mathematics gap between the richest and the poorest primary schools in the six provinces roughly halved between the 2018 and 2022 cohorts, from 0.48 to 0.23 SD. On the common scale, it did not narrow: it was 0.72 SD in 2018 and 0.77 in 2022. Each cohort is calibrated on its own scale; Section 4.6 shows why a change in the spread of either cohort’s marks could not produce this.

Calibration re-ranks most primary schools, and within each cohort the gap between the richest and the poorest schools is wider on the common scale than on raw marks. Ranked once by their students’ mean raw mark and again by their calibrated level of learning, 72.5% of schools land in a different quintile of learning. As assessments are hardest at the richest schools, a production function relating learning to school characteristics recovers about a third of the poverty gradient when it is estimated on raw marks. The assessments of no-fee schools, which the state bars from charging fees and which serve poorer communities, are 0.47 SD easier than those of fee-charging schools, the strongest association among the school characteristics I observe; at Grade 12, school marks overshoot the school-leaving examination least in the richest schools (Van der Berg and Shepherd 2015).

As an external test, I show that calibrated scores predict the national school-leaving examination better than raw marks do. Because no assessment is common to every primary school, I use the first examination that students share: the National Senior Certificate (NSC), sat five years after primary school. Linking the NSC to a third cohort, the students who left primary school in 2020, I find that among students at the same secondary school, the calibrated score explains about a quarter more of the variation in NSC results than the raw mark does.

Easier assessments do not buy a place at a higher-performing secondary school: at the same mark, students from primary schools with easier assessments enter secondary schools with lower-achieving classmates. A student whose primary school’s assessments are one SD harder across primary schools joins classmates whose mean mark is 0.165 SD higher, holding her own mark fixed; her primary school’s raw average, the number parents can see, predicts nothing about them (0.02), and the wedge predicts the published NSC results of the school she enters about three times as strongly as that average does. Below the leaving examination no results are published for schools, so students are sorted on a quantity no one could observe, an unusual margin for the school-choice literature, which measures responses to published scores (Hastings and Weinstein 2008; Abdulkadiroğlu et al. 2020).

With the calibrated score, descriptive value added becomes estimable for each of 1,135 secondary schools in the six provinces: how far each school’s students finish on the NSC above students who entered with the same calibrated score. Value added needs a comparable intake: on raw marks, a secondary school fed

by primary schools with harder assessments looks as though it taught more. Two in five secondary schools change their quintile of value added when it is estimated with the calibrated intake score instead of raw marks. The calibrated estimates meet the reliability standard for descriptive value-added research, split-half 0.86 against the 0.80 benchmark (Nunnally 1978, 245; Lance et al. 2006, 206); causal value added would also require assignment to secondary schools that is as good as random.

To my knowledge, this is the first paper to use the two-way fixed-effects design to put school-set assessments on a common scale where neither stage has a common test or a shared grader. Nascent work in the 1960s attempted to predict success at college by moderating schools' grades through the students who moved from secondary schools to tertiary institutions (Bloom and Peters 1961; formalised by Tucker 1963). That work did not use modern estimators such as two-way fixed effects, and has seen little subsequent development without a common test. In the education literature, the two-way fixed-effects estimator has separated teacher and school effects from student composition (Mansfield 2015; Kramarz et al. 2015), but there an external standardised test puts every student on one scale before estimation, so the test does the comparing and the estimator does the attribution. Two contemporaneous papers model the grades themselves: within one school, students who take several courses at once link those courses' grades (Yang 2026; Barker 2026), and across schools Barker (2026) bridges them with standardised tests.

The design closest to mine (Liu et al. 2025) also anchors middle-school effects on the secondary schools their students enter next, but a state test has already put their students on a common scale. In my setting, no schools at either stage share an assessment, so the estimator itself does the comparing across schools.

Researchers measuring school quality in low- and middle-income systems typically use purpose-built tests fielded to samples of schools or children (Singh 2015; Araujo et al. 2016; Bau and Das 2020; Andrabi et al. 2011; Bold et al. 2017). These samples have low coverage, whereas administrative marks cover every school but share no scale.

Central examination is the usual route to comparable grades. In Germany, an additional standard deviation of school-leaving grades earns 6% where the examination is central, against under 2% where the school sets it (Schwerdt and Woessmann 2017): employers pay more for grades they can compare. Where schools grade their own work, grading standards differ and matter (Himmler and Schwager 2013; Figlio and Lucas 2004). Grade inflation and manipulation have been studied as strategic behaviour at particular grade cut-offs (Diamond and Persson 2016; Dee et al. 2019). The wedge is a broader object: the difficulty of a school's assessments, carried in every mark it awards, far from these cut-offs as well as at them. Measuring it takes no central examination: the design needs no more data than education management information systems often already collect, namely the marks students are awarded and a student identifier that follows the student between schools.

Section 2 describes the setting and the data; Section 3 the model, the estimator, and identification, and why only primary schools get a wedge. Section 4 shows that the measure is precise, what calibration changes, in rankings and gaps, and which schools set easier assessments, and validates the measure against the NSC; Section 5 tests the estimates against simulated threats to identification and heaping at the promotion threshold. Section 6.1 reports the sorting, Section 6.2 value added, and Section 7 the uses and limits of the measure.

2 Setting and data

2.1 Marks set by schools

Grade 7 is the last year of primary school in South Africa. Assessments are set by each school, from drafting the assessment to the creation of the rubric to the final marking. As there is no moderation between schools to a common standard, nothing in principle means a 60 at one school is equivalent to a 60 at another. The only national examination in the system is the Grade 12 National Senior Certificate, and even its final mark is three-quarters examination and one quarter the school's own assessment. The school-based fraction is moderated by Umalusi, the education quality council, which rescales each school's marks in a subject so that their spread matches the same students' examination marks and their mean sits in a narrow band just above

their examination mean.⁶ Too much time passes between Grade 7 and the NSC for the NSC to anchor the measure, but not to test it (Section 4.7).

Grade 8 is the first year of secondary school. As with Grade 7, all assessments in secondary school below Grade 12 are school-set and school-graded. The content they cover is not. Every public school teaches one national curriculum, and in mathematics a single document sets out what is to be taught in Grades 7 to 9, each grade building on the one before (Department of Basic Education 2011a, foreword and §2.7).⁷ Schools differ in the papers they set and in how they mark them, not in what they are meant to teach. The Grade 8 Term 2 mark is used because this is the term in which students simultaneously sit a grade-wide assessment. The final recorded mark used in this paper is a composite of this simultaneous end-of-term assessment and continuous assessment tasks common to all students. The choice of Grade 8 term leaves the ordering of primary schools by wedge closely preserved, as Section 3 and Appendix D report. The Grade 7 measure is the Term 4 mark, the last report mark of primary school and the one that decides promotion. The two marks are about six months apart: Term 4 of Grade 7 closes in December, and Term 2 of Grade 8 in June (Department of Basic Education 2016, 2017, 2021, 2022a). Both are the student’s *mark* throughout this paper; *score* is reserved for the calibrated common scale of Section 3. By *learning* I mean the level a score measures: standing on what the shared Grade 8 assessments reward. For fairness between students, almost all schools set the same assessments for all students. This ensures the within-school comparability on which the estimator relies.

The calibration substitutes for the lack of nationally comparable assessments below the secondary leaving examination. No national assessment below the National Senior Certificate (NSC) reports results school by school, while the NSC’s results are published for every school each year (Department of Basic Education 2025). South Africa attempted to implement national assessments of every student in Grades 1 to 6 and 9, the Annual National Assessments, which were discontinued for several reasons, including union resistance (Department of Basic Education 2020a, 49–50; SADTU, NAPTOSA, SAOU, NATU and PEU 2015). South Africa does participate in international large-scale assessments, but these are sample-based, so they cannot create the school-level resolution education systems often require for effective management. National assessments expose education systems to accountability, which leads to their undersupply (World Bank 2018). Where political-economy considerations prevent comparable national assessment, a method that recovers comparability using only the transition from primary to secondary (marks that education management information systems often already track) can create some of the critical information comparable assessments provide.

2.2 The promotion threshold

Schools face systemic incentives that push marks in predictable directions. Promotion policy sets minimum marks by subject. In Grades 7 to 9 the requirement in mathematics is 40% (“moderate achievement”), while 50% (“adequate achievement”) is the requirement in the home language and 30% (“elementary achievement”) the floor in two other subjects.⁸ Schools face pressure to get students across these thresholds from parents, departmental accountability metrics (North West Department of Education 2024, 177, 286) and simply the cost of retaining a large number of students. The result is threshold pushing: inflating marks upwards across the nearest threshold.

Threshold pushing is evident in the data, as found by van der Berg et al. (2024). The marks heap where the policy binds: 40% is the mathematics requirement, and neither 30% nor 50% applies to mathematics. The pushing is tied to promotion: the Term 4 mark, which decides promotion, heaps at 40 more strongly than the Term 1 mark does up to Grade 9, and at Grade 12, whose promotion runs through the National Senior Certificate, the Term 4 excess disappears (Section 5.2).

Figure 1 shows threshold pushing in the Grade 7 mathematics data, and it sits almost entirely on one cut-off.

⁶Umalusi (2016), §9.5.

⁷The 2022 cohort’s two years ran on teaching plans revised after COVID-19, again the same for every school (Department of Basic Education 2020b, 2022b).

⁸Department of Basic Education (2011b), §21(1) and Table 17. The Senior Phase requirements are the same in the consolidated text of 5 February 2021, so they cover both cohorts.

Across the six provinces, marks are two to seven times as frequent just above 40 as just below it, and the step appears in every province and every year. It is already there in the first term of the Grade 7 year, and it is largest at Term 4. Grade 8 marks show a smaller step.

This threshold pushing is not classical measurement error with mean zero and does not average away in large samples: it moves marks only upwards, only at the thresholds, and depends on the student’s true mark, properties Section 5 takes up.

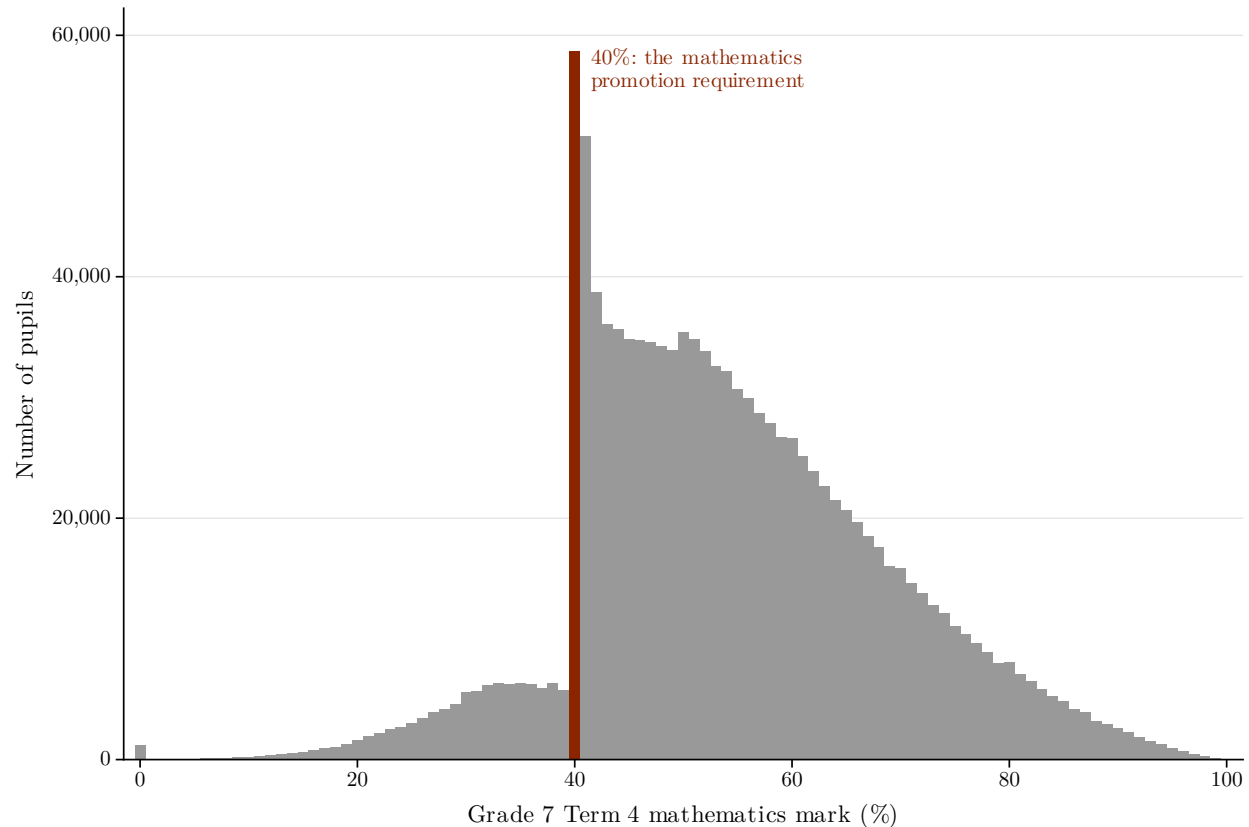


Figure 1: **Marks pile up at 40%, the mathematics promotion requirement.**

Note. Integer-bin histogram of Grade 7 Term 4 mathematics marks, 2018 and 2022 cohorts pooled. Only the highlighted bar at 40% carries a spike, the signature of school-set marks responding to the promotion rule rather than tracking learning alone; 30% and 50%, which promotion requires of other subjects but not of mathematics, carry little excess. The small residual at a mark of zero is not a promotion threshold. Students below 40 link to a Grade 8 record less often, so the jump at 40 is somewhat larger here than among all Grade 7 records. **Sample:** the six provinces, students linked to a Grade 8 record from primaries with at least ten linked students, before the restriction to the largest connected set ($N = 1,198,933$).

2.3 The move to secondary school

Almost all primary schools end at Grade 7. Combined schools, which teach both Grade 7 and Grade 8, are the exception. The move to Grade 8 transfers most students to a new institution and scatters them widely from their primary school peers. Across the six provinces, 86% of students in the estimation sample change school between Grade 7 and Grade 8 in 2018 and 92% in 2022; the rest stay on at a combined school.

In the 2018 mathematics sample of all six provinces, the median Grade 7 feeder, the primary school that sends students on, disperses its students across 6 distinct Grade 8 destinations (mean 7.7). The median secondary school receives students from 8 distinct feeder schools (mean 12.0). The network is dense and its

right tails are long (Table 2, Panels B–C). The destination is a secondary school’s Grade 8 in a given year; the records carry no class identifier. Each destination that receives a primary school’s leavers is another institution in which to compare that school against other schools’ students, so the more widely its leavers spread, the more comparisons enter. How evenly a primary school’s leavers spread across destinations is measured by a Herfindahl index: the sum of the squared shares of a primary school’s leavers entering each Grade 8 destination, equal to 1 when all its leavers enter one school and falling towards 0 as they spread over many schools in similar numbers. Gauteng’s leavers are the most evenly spread of the six provinces, with a median index across its primary schools of 0.26 in both cohorts, against 0.44 to 0.61 in the other five provinces. The basic nature of the mixing structure is visualised in Figure 2. Traced through an illustrative cluster of Grade 7 students, the flows fan out and cross repeatedly, so feeders that never share a Grade 7 assessment are joined and destinations that never share a Grade 8 assessment are likewise joined.

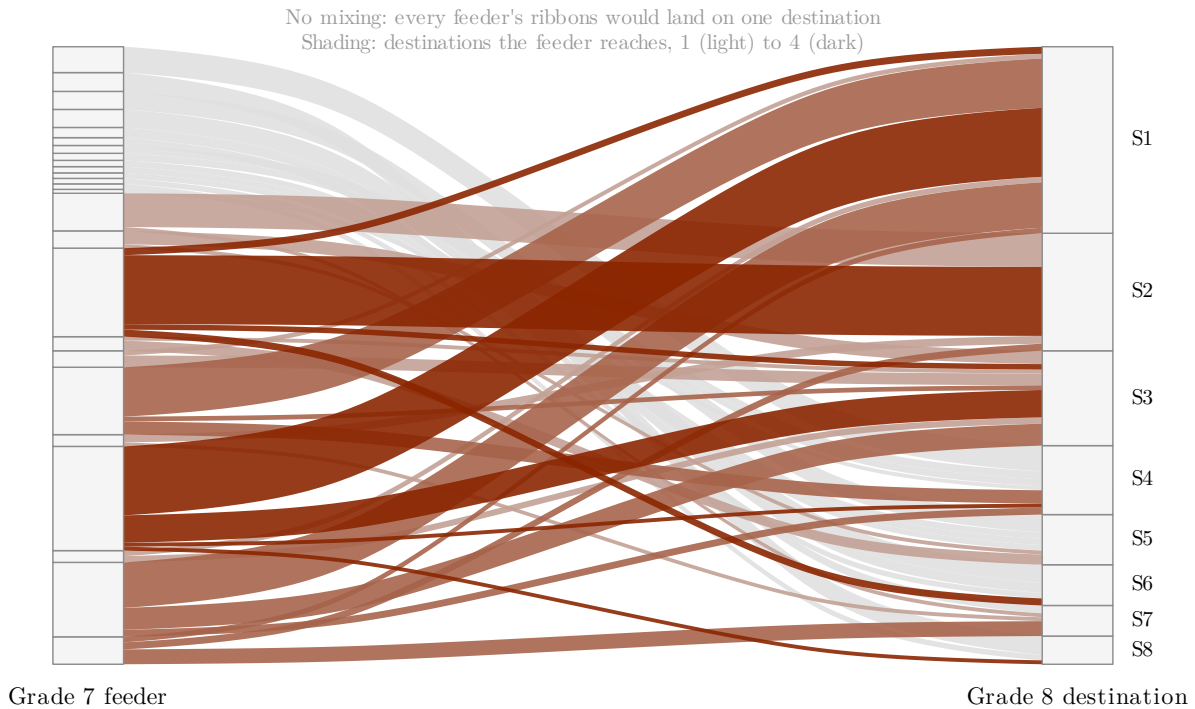


Figure 2: **Many feeders split their students across destinations, and each destination draws from several feeders.**

Note. Each ribbon is one flow of students from a Grade 7 feeder (left) to a Grade 8 destination (right), its width the number of students. Ribbon shade is the number of distinct destinations the feeder reaches, from light (one) to dark (4). The cluster is one connected component of the 2022 network, with 25 feeders, 8 destinations and 788 students. The cluster depicted is only illustrative: it is chosen because it is small enough to draw, so its widest feeder reaches only 4 destinations, against 7 for the median feeder in the full 2022 sample, and flows of fewer than 5 students are dropped. **Sample:** the six provinces, 2022 cohort, all subjects, one connected cluster.

What makes Grade 8 usable for comparison is not how it is set, which is identical to Grade 7, but the mixing of students from different sending schools into an environment where the same assessment is sat. Feeders are compared only through the set of assessments their leavers sit in common at a given secondary school. This many-to-many mixing creates the identifying variation this paper exploits. Students graded on assessments set and graded in different primary schools sit the same assessment at each Grade 8 destination; the relative standing of any two feeders becomes observable when those feeders mix their students in the same Grade 8 destination. As the network is dense, with each feeder reaching many destinations and each destination

drawing from many feeders, these pairwise comparisons can be chained together by the mutual overlapping of receiving and sending schools’ students, creating a single connected set. Without this mixing, the feeder schools would form disconnected islands, with each island incomparable to other islands.

2.4 Data and samples

Student-level records come from SA-SAMS, the administration system in which South African schools record their students’ marks, for two cohorts four years apart: Grade 7 in 2018 (entering Grade 8 in 2019) and Grade 7 in 2022 (entering Grade 8 in 2023).⁹ The extract comes from the Data Driven Districts programme (van der Berg et al. 2024). Mathematics is the focus of this study because the English First Additional Language calibration failed the reliability screen ([Appendix C](#)). The extract covers six of the nine provinces: the Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, and North West. The Western Cape keeps its schools’ records in its own system rather than SA-SAMS (Western Cape Provincial Parliament 2022). SA-SAMS records each student’s date of birth, sex, home language and population group, and the department’s master list of schools gives each school’s poverty quintile, fee status, district and rural status. The quintile ranks public schools by the poverty of the community around them, from 1, the poorest, to 5; schools in quintiles 1 to 3, and some others a province names, charge no fees (Department of Education 2006; Department of Basic Education 2014).

These cohorts straddle COVID-19 disruptions. Neither cohort’s move from primary to secondary school was disrupted, but a large disruption fell between the two cohorts. The 2022 cohort’s Grade 7 year still ran under the pandemic’s assessment rules, which gave school-based assessment through the year 80% of the promotion mark, against 40% in 2018 (Department of Basic Education 2011b; Eastern Cape Department of Education 2022), so its Term 4 mark is in effect a year mark: within a school, among students with all four Grade 7 term marks, a student’s Term 1 to Term 3 marks explain 94% of its variance, against 71% in 2018. I use a third cohort for the results that follow students to the school-leaving examination. The student-level test against the school-leaving examination in [Section 4.7](#) and the value added of [Section 6.2](#) use the Grade 7 cohort of 2020, calibrated separately under the same specification, because it is the only cohort I can follow to the examination student by student. Of that cohort, 394,353 students entered a secondary school in the six provinces with a calibrated intake score. For the 2020 cohort, both Grade 7 and Grade 8 experienced pandemic disruptions, a limitation [Section 4.7](#) discusses.

Grade 7 and Grade 8 records were linked within the same province using a two-pass algorithm, the first pass matching on the learner identifier and the second on date of birth, sex, home language, and population group, where that combination uniquely identified one student in both Grade 7 and Grade 8. [Table 1](#) traces the resulting construction funnel: of 796,928 (2018) and 866,679 (2022) Grade 7 enrolments in SA-SAMS, 88.9% and 92.8% carry a valid Term 4 mathematics mark, of which 76.8% and 83.8% link to a Grade 8 record with a valid Term 2 mark. After restricting the sample to the largest connected set within each province-cohort and to a minimum of ten linked students from each primary school, the estimation samples are 532,295 (2018) and 665,019 (2022) students. No minimum is set on the number of students a feeder sends to any one destination, so every student in the sample counts towards the wedge of the feeder that graded them ([Appendix S.3](#)).

There are three main samples used throughout the paper, clearly declared for each exhibit. The **national sample** is all six provinces pooled, and is typically used only to answer population-level questions, such as this section’s funnel and the balance check in [Appendix A](#). The **Gauteng core**, 106,514 students in 2018 and 135,852 in 2022, carries the main methodological and sensitivity analysis, because the results are not comparable across provinces, so one province must carry it, and Gauteng has the densest network ([Section 2.3](#)) and therefore the most identifying variation (most analyses are replicated for all provinces in the appendices). The **six-province panel** re-estimates the measure separately in each province, each on its

⁹The sample includes independent (private) schools. Leaving out every student whose primary or secondary school was independent in the year the student attended it (3.8% of linked students in 2018 and 3.7% in 2022) and refitting each province and cohort leaves the ordering of primary schools by wedge almost unchanged: the rank correlation with the full-sample wedge is at least 0.994 in every province and cohort, and the slope of the Grade 8 mark on the Grade 7 mark moves by at most 0.012. The spread of the wedge narrows most in Gauteng in 2018, to 0.92 of its full-sample standard deviation, because the independent primaries there set harder assessments than public primaries and differ more among themselves.

Table 1: **Two-thirds of 2018’s and three-quarters of 2022’s Grade 7 enrolments reach the sample.**

Step	2018 cohort		2022 cohort	
	N	%	N	%
Total Grade 7 enrolment in SA-SAMS	796,928		866,679	
With valid Grade 7 Term 4 maths mark	708,659	88.9	804,654	92.8
Successfully linked to a Grade 8 record	544,233	68.3	674,557	77.8
Not linked: repeated Grade 7 the next year	—		27,339	3.2
Final analysis sample	532,295	66.8	665,019	76.7

Note. Students remaining after each restriction, by cohort. Each percentage is of the cohort’s Grade 7 enrolment in SA-SAMS. The indented row shows the students who do not link because they did not meet the promotion requirements and repeated Grade 7 the following year; it is counted for 2022 only. The final sample keeps students in the largest connected set within their province and cohort, from primaries with at least ten linked students. **Sample:** the six provinces, mathematics, 2018 and 2022 cohorts.

own connected set. A fourth sample, the replication, is a different exercise: the same specification, applied without change to a separate national extract of school-based-assessment marks from the Department of Basic Education, Terms 1 and 4 of 2021 to 2024 in all nine provinces ([Appendix D](#)).

Table 2 describes the sample: students’ marks and characteristics in Panel A, and in Panels B and C how widely primary and secondary schools mix. The original marks run from 0 to 100; in the SD units I use, one SD is roughly 14 marks in Grade 7 and 18 in Grade 8 (the standard deviations of the two marks pooled over the six provinces, in both cohorts).

Table 2: **Schools in the sample mix widely across the move to Grade 8.**

Variable	2018 Cohort			2022 Cohort		
	Mean	SD	Mdn	Mean	SD	Mdn
<i>Panel A: Student characteristics</i>						
Grade 7 Term 4 maths mark	50.9	13.9	49	56.8	14	56
Grade 8 Term 2 maths mark	34.9	17.7	34	40.9	18.2	41
Over-age for grade (%)	32.1	46.7	0	29.6	45.7	0
Poverty quintile (1–5)	2.6	1.3	3	2.5	1.3	2
<i>Panel B: Grade 7 feeder schools</i>						
Students per feeder school	48.3	37.2	36	55.8	43.6	42
Grade 8 destinations per school	7.7	6.1	6	8.6	6.8	7
<i>Panel C: Grade 8 schools</i>						
Grade 7 feeder schools per secondary school	12	12.4	8	15.5	14.4	11

Note. Panel A reports student-level marks and demographics; Panels B–C report the feeder-to-destination network, computed on distinct schools within each cohort. Over-age means turning 14 or older in the Grade 7 year. A school’s poverty quintile is its quintile in the year its students were in Grade 7. Panel C counts secondary schools, not classes. The table describes linked students from primaries with at least ten linked students, before the restriction to the largest connected set, slightly more than the final sample. **Sample:** the six provinces, mathematics, 2018 and 2022 cohorts. N = 533,419 (2018), 665,514 (2022).

The most critical level of attrition would occur if a large number of feeder schools were not linked to the largest set of connected schools through the students they send on, creating islands that cannot be placed on the common scale. Only 103 primary schools, 0.45 per cent of the total and holding 0.17 per cent of students, fall outside the largest connected set of their province and cohort.

Sample attrition raises selection concerns, particularly that lower-performing students are held back. This would bias the results if held-back students were concentrated in schools with easier or harder assessments. I find that linkage is close to orthogonal to the wedges themselves. Across 22,868 school-years, measured within each province and cohort, the correlation between a school’s wedge and linkage rate is -0.034 (i.e. linkage explains under one per cent of the variance in the wedges). An inverse-linkage-probability weighting (IPW) estimator shifts the mean wedge by 0.004 SD, substantially below the precision of the estimates themselves. The full diagnostics, including the balance table, the linkage–mark mechanism, the IPW specification, connected-set coverage by province, and the students who do not progress to Grade 8, are reported in [Appendix A](#).

3 Model, estimator, and identification

Two students leave two different primary schools, A and B, with the same Grade 7 mathematics mark, 65, and enter the same secondary school. At the secondary school, they sit the same Grade 8 assessments, and primary school A’s student gets 72 while B’s gets 58 (Figure 3, Panel A). If 65 had meant the same amount of learning at both primary schools, the two would achieve the same mark in expectation in Grade 8. The 14-point difference measures how much harder A’s assessments were than B’s: A set harder questions, marked them more harshly, completed more of the curriculum, etc. This difference is then revealed in the Grade 8 marks. Every secondary school that receives students from both A and B repeats the comparison.

Just as convergence from two primary schools to one secondary creates the comparison of those primary schools, divergence from one primary school to two secondary schools allows those secondary schools to be compared through students who got the same mark on the primary school’s assessments. Take B’s student above, who got 58 at Secondary 1, and another of B’s leavers with the same 65, who went to Secondary 2 and got 53. B graded both alike, so the 5 points are the secondaries’. Comparisons through a shared secondary school rank primary schools, and comparisons through a shared primary school rank secondary schools. Chained together, they put almost every primary school in a province on one scale (Panel B). The regression in this section makes all these comparisons at once. For each primary school it estimates a single parameter, which I call its difficulty wedge, that best fits all of the comparisons involving its students. The remainder of this section sets out how the estimator works: Section 3.1 presents a model of marks and learning to show what each coefficient in the regression captures, Section 3.2 reports the empirical estimates and the extent to which primary schools’ assessments differ, and Section 3.3 discusses the assumptions required for a gap like the 14-point difference above to reflect only the difficulty of a school’s assessments, including how strictly they are marked. Section 3.4 turns the estimates into each student’s calibrated score and shows which schools set easier assessments.

Timing explains why I can isolate difficulty for primary schools, but not for secondary schools. When students from Primary A and Primary B both leave with a mark of 65 and attend Secondary 1, they experience the same Grade 8 environment and sit the same June assessments, which means that the 14-point gap between them reflects the difference between Primary A’s and Primary B’s assessments. By contrast, when two students leave Primary B with 65 and attend different secondary schools, they spend the first half of Grade 8 in different classrooms before sitting different June assessments, so the 5-point gap between them cannot separate what each secondary school taught before June from the difficulty of its assessments. Although the regression estimates a fixed effect for each secondary school, which I refer to as its “destination effect”, this single parameter combines the school’s first half-year of teaching with the difficulty of its assessments. Only primary schools get a wedge.

3.1 The model and the regression

Take one student, i , who leaves primary school j at the end of Grade 7 and enters the Grade 8 grade of secondary school k , the grade rather than the school because every student in the grade sits the same assessments (Section 2). Call what she knows at the end of Grade 7 her learning, a_{ij}^7 , and what she knows by the middle of Grade 8 a_{ijk}^8 , both measured, as in Section 2, by what the shared Grade 8 assessments reward. Neither of these is observed. Instead, what is observed is the mark each school gives her, g_{ij}^7 and g_{ijk}^8 , and a school's mark differs from her learning in two ways: by the shift implied by the marks a school gives everyone, and by noise particular to her,

$$a_{ij}^7 = g_{ij}^7 + s_j + \varepsilon_{ij}^7, \quad a_{ijk}^8 = g_{ijk}^8 + d_k + \varepsilon_{ijk}^8. \quad (1)$$

The shift s_j describes how hard primary school j 's assessments are: positive when the school's marks are below its students' learning. Difficulty includes how assessments are set and how they are marked. The shift d_k is the same sort of object as s_j , but now for the Grade 8 grade of secondary school k . In the opening example, A sets harder assessments than B. The two noise terms are zero in expectation and not related to each other or to learning. Between primary and secondary, learning is assumed to increase or decay at the same expected rate for all students,

$$a_{ijk}^8 = c + \lambda a_{ij}^7 + \nu_{ijk},$$

where $\lambda \in (0, 1]$ is the persistence parameter, c is the average gain over the transition, and ν_{ijk} is the student's own departure from it. At $\lambda = 1$ a student retains all learning between primary and secondary; at $\lambda < 1$ part of it fades.

Substitute the two equations into the third and only marks remain:

$$g_{ijk}^8 = \beta_0 + \beta_1 g_{ij}^7 + \alpha_j + \gamma_k + u_{ijk}, \quad (2)$$

with $\beta_0 = c$, $\beta_1 = \lambda$, $\alpha_j = \lambda s_j$ and $\gamma_k = -d_k$, and an error u_{ijk} that collects the two noise terms and also how much more or less this student learned than others who left Grade 7 with the same learning (not marks) as she did (Appendix O gives the steps).

This is the model which is finally estimated: a student's Grade 8 mark is regressed on her Grade 7 mark, with fixed effects for the primary school she left and the Grade 8 of the secondary school she entered. The overall effect of the primary school is how hard its assessments are, scaled by persistence. The primary effect, $\hat{\alpha}_j$, is what this paper denotes as the school's wedge, which is positive where a school's assessments are harder and negative where they are easier. The destination effect, γ_k , absorbs the Grade 8 assessment, including how difficult it is, how it is marked, and anything else common to students in that grade. Since $\gamma_k = -d_k$, its sign is reversed: a negative destination effect describes a school that gives lower Grade 8 marks for the same Grade 7 mark. The two shifts enter differently only because the primary school's shift is inside the regressor and is scaled by persistence, while the secondary school's shift is inside the outcome, entering without this scaling.

If there were no noise in the marks or decay in learning between primary and secondary school, the coefficient on the Grade 7 mark would be one. Instead, it is 0.60 in 2018 and 0.59 in 2022 (Table 3). The coefficient is pulled beneath one because learning fades across the transition, so that $\lambda < 1$, and because measurement noise attenuates the slope by the mark's reliability, ρ , which is the share of its variance that is signal after both schools' fixed effects are removed:

$$\hat{\beta}_1 \xrightarrow{p} \lambda \rho. \quad (3)$$

The regression cannot differentiate these two parameters. Any combination of λ and ρ whose product equals the slope fits the regression equally well; as each is at most one, learning is at least 0.59 persistent and the Grade 7 mark is at least 59% signal.

A second mark for the same student separates persistence from reliability. Students who take English First Additional Language, the English course for students whose home language is another language, also have a Grade 7 English mark. It comes from different assessments, so its noise is largely unrelated to the noise

in their mathematics mark, while it tracks learning at the same school. The slope on the part of the mathematics mark that the English mark predicts is free of the mathematics mark’s noise, so it measures persistence alone: about 0.76 in 2018 and 0.73 in 2022 ([Appendix N](#)). For comparison, over a school year in Pakistan, Andrabi et al. (2011) find that 70–80% of achievement persists when test scores are corrected for measurement error alone, as here, and a fifth to a half when students are also allowed to learn at different rates. Dividing the ordinary slope for the same students by that persistence gives the mark’s reliability: about 0.76 in 2018 and 0.78 in 2022. Noise and fade share the shortfall roughly equally, with fade a little the larger in 2022. Both marks heap at 40%, so a little of their noise is shared, which pulls the instrumented slope towards the ordinary one: if anything, these figures understate persistence and overstate reliability.

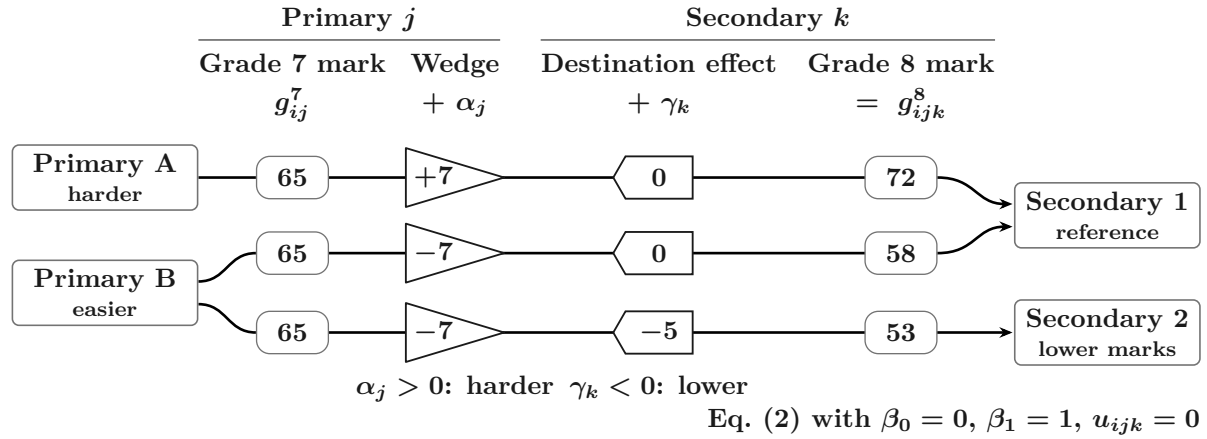
Because the Grade 7 mark sits on the right-hand side of the regression, its noise attenuates the slope, and through the slope it reaches the wedge:

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

up to a constant, where $\bar{a}_j^7$ is the mean learning of school j ’s leavers. The first term is difficulty, attenuated by the same factor as the slope. The second is intake: noise averages out of a school’s mean mark, so the attenuated slope underpredicts the leavers of a school whose students know more, and that school’s assessments look a little harder than they are. Recomputing the wedges with the slope set at persistence, measured with the English mark, removes both distortions. The spread of the wedges widens from 0.471 to 0.554 within-province standard deviations (SD) of students’ Grade 8 marks in 2018 and from 0.468 to 0.527 in 2022. The gap between no-fee and fee-charging primary schools widens from 0.47 to 0.51 SD; [Section 4.5](#) takes it up. However, the two sets of wedges are correlated at 0.994 in 2018 and 0.995 in 2022, which means that the ranking of primary schools is almost unchanged. Differences in student intake alone account for only 9–10% of the variance in the wedge and 13–15% of the gap between no-fee and fee-charging schools. Because the main analysis uses the ordinary least squares estimates throughout the paper, both the dispersion of the wedges and the no-fee gap reported below are, if anything, conservative ([Appendix N](#)).

3.2 Estimation

(A) Two comparisons



(B) Chained onto one scale

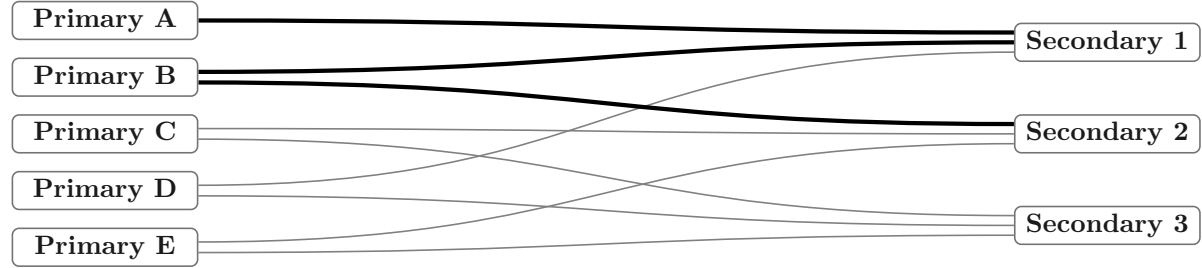


Figure 3: **Same mark in, different marks out: the gap between primary schools measures difficulty.**

Notes: The marks are illustrative. Panel A: each row is one student. The sum of her Grade 7 mark, her primary's wedge, and her Grade 8 grade's destination effect gives her Grade 8 mark, as Equation 2 with an intercept of zero, a slope of one, and no noise. The wedges are centred on zero, so each is read against the average of A and B; Secondary 1's destination effect of zero follows from that centring and the zero intercept. At Secondary 1, A's and B's students share the destination effect, so their 14-point gap is the gap in wedges; B's two students share the wedge, so their 5-point gap is the gap in destination effects. Panel B: each line is the students moving between two schools; bold lines are Panel A's moves. In the regression, wedges are in SD and centred within each province and cohort.

Figure 3 illustrates the two comparisons from the opening example (Panel A) and how student transitions chain them together across a network of schools (Panel B). The wedge is measured in Grade 8 mark units, where the comparison takes place, so B's wedge lies 14 points below A's. Because only a fraction λ of Grade 7 learning persists into Grade 8, the underlying Grade 7 gap is larger than the 14-point gap observed in Grade 8, which is why $\alpha_j = \lambda s_j$ in Equation 2. For simplicity, Figure 3 sets λ at one, so the two gaps are identical in the diagram. Because the data identify only relative differences between primary schools, centring the two schools around zero assigns a wedge of +7 to A and -7 to B. Crucially, this comparison does not require knowing how Secondary 1 sets or marks its own assessments, because both students face the same Grade 8 standard there. In the data, students disperse widely across secondary schools (Section 2.3), linking almost every primary school in a province into a single connected network. Across this network, $\hat{\alpha}_j$ captures the average extent to which primary school j 's leavers score higher or lower in Grade 8 than their Grade 7 marks would predict, conditional on the secondary schools they attend, while $\hat{\gamma}_k$ captures the extent to which students entering Grade 8 at secondary school k score higher or lower than predicted by their Grade 7 marks and primary schools.

In the estimates, as with the example, the wedges are centred on zero within each province-cohort sample.

Each $\hat{\alpha}_j$ is indexed against the average primary school and never against a secondary school.¹⁰ In every province and cohort, the largest connected set contains at least 98.8% of primary schools (Appendix S.2), so the scale covers effectively every school. The model is estimated within provinces, because the records are linked within each province (Section 2). Little is lost: the links between provinces are thin. In a separate match, 1.9% of linked students cross a provincial border, at 1.14 students per feeder-destination pair against 8.2 within province.¹¹

I estimate Equation 2 by ordinary least squares with high-dimensional fixed effects (Bergé et al. 2026) on marks standardised within province and cohort. Section 6.1 standardises differently, each signal and outcome by its own spread over the moving students of a cohort, and says so where it does. The choice of Grade 8 term barely matters: net of estimation noise, wedges estimated from any other Grade 8 term correlate with the Term 2 wedges at 0.99 or more in both cohorts (Appendix D). Table 3 reports the estimates. The slope on the Grade 7 mark is 0.601 (standard error 0.010, two-way clustered by primary and secondary school) in 2018 and 0.592 (0.007) in 2022. The wedges spread widely. Their standard deviation across schools is 0.471 SD in 2018 and 0.468 in 2022, and schools at the 10th and 90th percentiles stand 1.13 SD apart on average across the twelve province-cohort samples. Two primary schools’ leavers, with the same Grade 7 mark, can land more than a full SD of Grade 8 marks apart. Gauteng is not unusual; the other provinces show much the same spread (Section 4.1, Appendix D). The destination effects spread more: their standard deviation across Grade 8 grades is 0.578 SD in 2018 and 0.518 in 2022 (Appendix F).

The paper reports results in two units, because each grade’s marks are standardised separately. The wedges, the destination effects and the calibrated levels are in standard deviations of Grade 8 marks, the marks in which the regression compares schools. A school’s raw mean mark, and a student’s own mark, are in standard deviations of Grade 7 marks. A gap in one unit is not the same size as a gap in the other, because learning fades between the two grades. About three-quarters of a difference in learning at the end of Grade 7 is still there in the middle of Grade 8 (Section 3.1). If one school’s leavers are 0.4 SD ahead of another’s in what they know at the end of Grade 7, they will be about 0.3 SD ahead in Grade 8 marks at the same secondary school. Wherever the paper sets a raw gap beside a calibrated one, it first multiplies the raw gap by persistence for that province and cohort.

Table 3: **Wedges differ widely across schools; the Grade 7 mark’s slope is below one.**

	2018	2022
<i>Panel A. Specification</i>		
Students	106,514	135,852
Origin (Grade 7) schools	1,424	1,479
Destination (Grade 8) grade cells	901	839
Slope on the Grade 7 mark	0.601	0.592
Standard error (two-way clustered)	(0.010)	(0.007)
Adjusted R-squared	0.678	0.609
Within R-squared	0.404	0.398
<i>Panel B. The wedge</i>		
Standard deviation	0.471	0.468
Median standard error	0.067	0.067
Different from the average school at 5% (%)	70.4	74.8

Notes: Ordinary least squares estimates of Equation 2 in each cohort, with fixed effects for the primary school and the Grade 8 grade cell. Marks are standardised within province and cohort, so the slope is in within-province SD of the Grade 8 mark per SD of the Grade 7 mark, and Panel B’s first two rows are in SD. The slope’s standard error is two-way clustered by primary

¹⁰A common scale is estimable for Grade 7 schools or for Grade 8 schools, but not one that puts both on the same scale. The destination effects, γ_k , are the common reference and do not enter as offsets comparable to the primary school wedge, so the wedge is only relative to the other origins.

¹¹The sample covers six of the country’s nine provinces, so a move from a sampled province into an unsampled province is recorded as attrition, with the true crossing rate likely a little higher than estimated.

and secondary school. Panel B’s per-school standard errors leave out the estimation error in the destination effects; Table 4 gives the leave-out noise share. The last row is the share of schools whose wedge differs at 5% from the average school of their province and cohort. **Sample:** Gauteng core, mathematics, 2018 and 2022 cohorts; N in the first row.

The wedge is not a ranking of students. A’s wedge of +7 is one number for the school; as the number of students grows, the grading noise averages out of it. The wedge is also not a causal effect of the school on learning. Nor does it reward a school for the students it admits. The comparison is between students with the same Grade 7 mark, so a stronger intake raises a school’s marks, not its wedge, apart from the small intake part that noise adds (Section 3.1) and any learning not captured by the Grade 7 mark (Section 3.3).

3.3 Identification

The regression reads the 14 points between A’s and B’s students as difficulty. That reading assumes that two students who leave primary school knowing the same amount earn, on average, the same Grade 8 mark at a given secondary school, whichever primary they left, and that their Grade 7 marks differ, beyond noise, only by difficulty. These model conditions (Kline 2024, 126) imply that a student’s learning gain or decay between primary and secondary is independent of the primary she left, and that the secondary school she sorts into does not depend on the speed at which her learning would improve.

The model is the two-way fixed-effects model of Abowd, Kramarz, and Margolis (1999, AKM hereafter), with students as the workers and the two schools that mark a student as the two firms: the feeder in Grade 7 and the destination grade cell in Grade 8. Where this model recovers a firm’s pay premium in the labour literature, here it recovers how hard a primary school’s assessments are. The main departure is that a student’s own effect is not a fixed effect but her Grade 7 mark, a noisy number with a slope, which is why $\hat{\beta}_1$ estimates $\lambda\rho$ and why noise in that mark reaches the wedge (Section 3.1). There are two comparative strengths of this setting over the labour literature: moving to a secondary school is mandatory for promoted students, and it happens simultaneously for all students. This fixes whether and when a student moves, though, not where (Kramarz et al. 2015, 577). Students at combined schools, 7.9% and 5.0% of the 2018 and 2022 samples, change grader without changing school, which is all the design needs of a move. Table 4 reports the standard checks of the two-way fixed-effects literature on the wedge (Kline 2024, 140–45; Bonhomme et al. 2026, 16–17), and Appendix S gives their detail. The departures are small, except one: whether a school’s students are all shifted by the same wedge. Below, each check is given with the assumption being tested, and Appendix O.9 lists every condition formally.

The first assumption is that, for a given level of learning (not marks) when a student exits primary school, which school she left makes no further difference to her marks. This is AKM’s exclusion restriction (Kline 2024, 140; Bonhomme et al. 2026, 20). In the model’s terms, given her learning, a student’s departure from the average gain does not depend on the school she left:

$$E[\nu_{ijk} \mid a_{ij}^7, j] = 0.$$

The data cannot test this directly. Going back to the first comparison of Figure 3: from the same Grade 7 mark, primary school A’s student gets 14 points more than B’s in Grade 8, and the regression reads the gap as A setting harder assessments. It would read the gap the same way if A’s students simply learned more after leaving, because A had taught them to work on their own, say. A single Grade 8 mark shows the same gap under either story, so I maintain the assumption.¹²

Whether A’s students learn faster after leaving can be checked in part. Because Grade 8 has four terms of marks, I can see whether the trend within Grade 8, which plausibly proxies for the trend between Grade 7 and 8, affects the estimates. Between Terms 2 and 4 of Grade 8, the marks of students from primary schools with harder assessments increase by 0.033 SD compared with those of students in their secondary school from primary schools with easier ones, for each SD of difference between the two primaries’ wedges. Difficulty itself produces a gain like this. Term 4 marks follow the Grade 7 mark about a tenth more closely than Term 2 marks do, so a primary school’s wedge shows up about a tenth larger by Term 4: in the example, A’s 14

¹²Di Addario et al. (2023) can test this assumption in the wage context, as the firm a worker left enters their equation as its own term, whereas here difficulty and the primary school’s lasting effect on learning enter as the same parameter, rather than as a learner fixed effect per se.

points would become about 15 with no extra learning at all. Difficulty alone predicts a gain of 0.042 SD, more than the 0.033 observed, so students from primaries with harder assessments learn no faster during Grade 8 than their classmates ([Appendix H.8](#)). Two conditions cannot be tested. The summer holiday between Grade 7 and Grade 8 has no assessment; I try to bound the effect of decay over it in the same appendix. And something a student learned in primary school that was not tested in Grade 7 might show up from the first Grade 8 term, just as difficulty does, so nothing here tells the two apart. Such learning would have to be sufficiently orthogonal to students’ Grade 7 marks, and differ between primary schools, to bias the results.

The same limit is true the other way. Unlike test equating, statistical moderation uses an anchor assessment to rescale each school’s marks. It also interprets any gap between the anchor and the original test as a school parameter. The reading is sound only if the school’s marks and the anchor measure the same thing (Masters 1988) ([Appendix Q](#)). Here the shared Grade 8 assessments are the anchor. The regression compares only averages and interprets the gap between a school’s marks and its students’ marks on those assessments as difficulty. Although the national curriculum fixes what schools are supposed to test, how much of that curriculum schools test is likely underdetermined. Had B taught only part of the Grade 7 curriculum, or taught it at a lower level, and marked fairly on what it taught, its student with 65 would still get 58 at Secondary 1, and the regression would still read B’s assessments as 14 points easier. Testing a narrower share of the curriculum is part of what makes a school’s assessments easier, and the estimator cannot separate incomplete curriculum coverage from lenient marking. Under either interpretation, a mark of 65 from Primary B represents less learning on the common scale, even though the wedge alone cannot determine whether Primary B needs to cover more of the curriculum or mark more strictly.

The second assumption, AKM’s exogenous mobility, is that among students from the same primary school with the same Grade 7 mark, which secondary a student enters is unrelated to how much she learns over the transition and to any learning her mark missed. Formally, given her Grade 7 mark and her primary school, the secondary she enters is independent of her error in Equation 2 ([Appendix O.9](#)):

$$k(i) \perp u_{ijk} \mid g_{ij}^7, j.$$

In the opening example, the 5-point difference between Primary B’s two leavers identifies the difference between the two secondary schools’ destination effects only if the student who entered Secondary 2 was not already on a slower learning trajectory. This assumption does not rule out school choice in general, as sorting on a student’s Grade 7 mark is controlled for directly by the regression slope. Rather, it rules out sorting across secondary schools among students from the same primary school who earned the same Grade 7 mark, whether based on their rate of improvement or on dimensions of learning that the Grade 7 mark failed to capture. Without a student fixed effect, the assumption asks more than AKM’s, which allows sorting on each person’s own effect (Card et al. 2013, 989); the student-fixed-effect version in [Appendix S.7](#) allows that sorting but imposes persistence of one.

Students do sort into destinations, but mainly on the Grade 7 mark the regression conditions on; sorting on gender and over-age is weaker, though beyond chance ([Appendix F](#)). Sorting on how fast students improve would show before the move, because Grade 7 has four terms of marks: had B’s leaver bound for Secondary 2 been improving more slowly, her marks would already have risen less over Grade 7 than those of B’s leaver bound for Secondary 1. In Grade 7 marks, students from one feeder bound for the top and for the bottom quarter of destinations gain almost the same over the four Grade 7 terms ([Table 4](#)): within 0.001 SD of each other with destinations ranked on split-half estimates, and within 0.012 SD with the other cohort’s, where the wedges themselves spread by 0.471 SD of Grade 8 marks in 2018. This is the pre-trend test of exogenous mobility (Card et al. 2013, 983–86). In 2018 students sort a little on transitory mark shocks, which leaves the order of the wedges almost unchanged ([Appendix S.4](#)).

Table 4: **Most two-way fixed-effects checks find small departures; a feeder’s wedge varies with the student’s mark.**

Assumption: the threat to it	Check (where reported)	Result	Reading
<i>Panel A. Standard two-way fixed-effects checks</i>			
Connection: schools fall into groups that no student moves between, so their wedges are not on one scale	Coverage of the largest connected set and of the leave-one-out connected set (Appendix S.2)	At least 98.8% of primary schools in every province-cohort cell; Gauteng 99.4% and 99.7%; the leave-one-out set keeps 1,421 of the 1,424 and 1,472 of the 1,479 connected primary schools	Connected
Connection: with few students on each link between a feeder and a destination, estimation noise inflates the spread of the wedges	Leave-out bias correction of the variance (Appendix S.3)	Noise is 5.4% and 5.5% of the wedge’s variance; SD 0.458 and 0.455 corrected, against 0.471 and 0.468	Small, not nil
Exogenous mobility: students bound for different destinations were already on different paths before the move	Marks of students who change school over the four Grade 7 terms, by destination, within the same feeder (Appendix S.4)	Pooled: split halves 0.001 (0.005); other cohort 0.012 (0.006); mean mark 0.027 (0.006)	Near flat by destination effect; a few hundredths of an SD by destination mean mark
Additivity: the two schools’ effects do not add: particular pairs of feeder and destination, or the change at the move, depart from their sum	Pair effects net of noise and mean residuals by wedge decile $\times$ destination decile; a slope per destination (Appendix S.5)	A separate effect per pair raises adjusted R^2 from 0.678 to 0.697 and from 0.609 to 0.629; net of noise, pair effects depart from the sum with an SD of 0.079 and 0.078 on links of two or more students; largest absolute cell mean 0.039 and 0.037; joint test $p = 0.020$ and 0.194 ; with a slope per destination, rank correlation 0.981 and 0.994	Small departures, rejected jointly in 2018; none at the move, net of composition
Exogenous mobility: where a student goes depends on something the model does not condition on	Within-feeder permutation test of destinations on observed traits (Appendix F)	Pooled: excess over chance, Grade 7 mark 16.9, gender 4.1, over-age 5.8 percentage points	Mainly on the conditioned mark

(continued)

Assumption: the threat to it	Check (where reported)	Result	Reading
Additivity: a feeder does not grade all its students by one shift	Mean residuals by the student's Grade 7 decile $\times$ wedge decile; a slope and shift per Grade 7 quintile (Appendix S.5)	Largest absolute cell mean, net of the Grade 7 margin shared by all feeders, 0.369 and 0.374 (low-mark students at the feeders with the easiest assessments); rank correlation with the baseline 0.997 and 0.999 after the quintile refit; with a slope per feeder, 0.979 and 0.991 at the average mark	Size varies with the mark; order barely moves under a slope per feeder
Additivity: a school grades some groups of its own students differently	Separate wedges by student group within the same schools (Appendix S.6)	Over-age on on-age, slope 0.72 and 0.76 against 0.94 and 0.94 from noise alone; girls on boys, 1.01 and 1.01 against 0.92 and 0.93	Narrower for over-age, by their flatter Grade 7 slope; slightly wider for girls, by their steeper slope and higher marks
The student's own effect: conditioning on the Grade 7 mark stands in poorly for each student's own effect	A student fixed effect from the eight term marks (Appendix S.7)	Rank correlation with the baseline wedge 0.876 and 0.949	A different object; similar order in 2022, less in 2018
Stability: a feeder's wedge drifts across cohorts, so one number per school misleads	Stationarity (Appendix G)	The wedge correlates 0.624 between the 2018 and 2022 cohorts, precision-weighted (1,280 schools)	Bears on use, not identification
<i>Panel B. Specific to school-set marks</i>			
Exclusion restriction: a feeder's effect on the Grade 8 mark is not all difficulty: the feeder taught skills its marks missed, or its students grow differently after the move	The exclusion restriction (Appendices D and H.8)	Pooled: 0.033 (0.002) per SD of the wedge over Grade 8 Terms 2–4, 0.022 (0.002) with the other cohort's wedge; Term 1 and Term 4 wedges correlate 0.902, precision-weighted (6,676 feeder schools, separate national extract)	Maintained; the Grade 8 gain is no larger than difficulty alone predicts

Notes: Each row names a threat to the wedge's identification, the test the two-way fixed-effects literature applies to it, the appendix that reports it, and the result. Panel B's row cannot be tested directly here: difficulty and any lasting effect of the feeder enter the same parameter. Magnitudes are in within-province SD unless a row names another unit, and row A5's are in SD of Grade 7 term marks; rank correlations are Spearman's, other correlations Pearson's. Cohorts are estimated separately, 2018 first, unless a row says pooled. Standard errors and the joint test are clustered by primary school. Each check uses its own sample, defined in Appendix S.0. **Sample:** Gauteng core, mathematics, 2018 and 2022 cohorts (106,514 and 135,852 students; 1,424 and 1,479 primary schools).

AKM's two remaining conditions are that the two effects add, one shift per school, the same for every student and every move; and that the schools are connected by movers. Connection holds for almost every school (Section 3.2). When only a small number of movers link schools to one another, sampling error inflates the estimated variance of the school fixed effects, which is the limited-mobility bias documented in sparse

labour-market panels (Andrews et al. 2008; Kline et al. 2020). Here the network is far thicker than in the labour research: by the leave-out correction, estimation noise is 5.4% and 5.5% of the wedge’s variance in 2018 and 2022 (Table 4), so the bias is small, though not nil.

The largest departure is from additivity within a feeder: a school does not shift all its students’ marks by one amount. Net of the Grade 7 margin shared by every feeder, low-mark students at the feeders with the easiest assessments land 0.369 SD above the model’s prediction in 2018, against a wedge standard deviation of 0.471 (Card et al. 2013, 995–96; Bonhomme et al. 2026, 19) (Appendix S.5). The size of B’s wedge, -7 in the example, therefore depends on which of its students one looks at. Its rank does not. A slope and shift per Grade 7 quintile leaves the order of feeders almost unchanged, and no alternative specification except the student fixed effect in 2018 reorders them appreciably (Appendix S). Across destinations the effects do add: a separate effect for every pair of feeder and destination adds little fit, though the pairs are jointly significant in 2018 (Table 4).

The cardinal size of a school’s wedge is more exposed than its rank. The placebo of Appendix H.2 shows how the size of $\hat{\alpha}_j$ rises with the part of the Grade 8 mark predicted by socioeconomic status, so the wedge still partly reflects selection. Part of that is the small intake part that noise adds (Section 3.1). Controls for over-age, gender and population group barely reorder the wedges (Section 4.2), and in the Monte Carlo simulations, sorting, thin or endogenous mobility, and attrition distort their size but leave their order largely intact (Section 5.1).

Two threats are particular to school-set marks. Marks pushed over the 40% promotion threshold (Section 2.2) are not symmetric noise, so no symmetric bounding exercise applies, including the bound on sorting bias in Appendix F. An aggressive smoothing of both mark distributions, a global polynomial that moves marks across the whole range rather than only out of the heap, makes no attempt to recover the true distribution: it changes the marks more than removing the heap would, which makes it a tougher test. It moves the slope little and leaves the order of the wedges in place (Section 5.3). At the destination the pushing follows the grade-wide promotion rule (Section 5.2), so it enters the destination effect. And a Grade 8 class teacher may mark differently from the teacher next door. Within a grade the classes sit the same grade-wide assessment and tasks common to all students (Section 2), so only the marking of them can differ; that is error, and it leaves the wedge unbiased so long as it is unrelated to where the students came from. There are no class identifiers in the data, so that condition is assumed.

3.4 The calibrated score, and the shape of the wedge

A student’s calibrated score is the place the regression puts her on the common scale: her Grade 7 mark, multiplied by the slope, plus her school’s wedge. The opening’s two students both have a Grade 7 mark of 65, but A’s student has a calibrated score 14 points above B’s, because A’s assessments are harder. A school’s calibrated learning level is the average of its students’ scores. One slope for every student fits the two ends of the Grade 7 range less well. Students with a Grade 7 mark below 30% land on average 0.15 SD higher in Grade 8 than their calibrated scores say, and students with 80% or more land 0.16 SD higher (2018). A school’s level averages over all its students, so it barely moves: levels computed with a separate slope for each of the seven achievement bands that schools report marks in correlate with these at 0.997 or more in every province and cohort. Averaging adds no new identification, since the regression has already placed every school on the common scale, but it defines the level at which the measure is used. To the first order of Equation 4 below, the level is the school’s mean Grade 8 mark net of destination effects: where its leavers land.

The wedge varies more across schools than the level does: a standard deviation of 0.471 against 0.418 SD in 2018 (0.468 against 0.446 in 2022). It is not poverty, fee status, or location under another name: together with every other school characteristic of Section 4.5, they explain 37.1% of the variance of the wedge averaged over the two cohorts. The wedge runs opposite to the raw Grade 7 mean, with a size-weighted correlation of -0.58 , averaged over the six provinces and both cohorts. This is in part arithmetic. To a first order, the bridge identity is

$$\hat{\alpha}_j \approx \bar{g}_j^8 - \hat{\beta}_1 \bar{g}_j^7 \quad (4)$$

which equates the wedge to a school’s mean Grade 8 mark, net of the destination effects, less a $\hat{\beta}_1$ fraction of its mean Grade 7 mark. Where two schools’ leavers do equally well in Grade 8, the school with the lower Grade 7 mark has the higher wedge because its assessments are harder. Among these schools, anything increasing with a school’s raw primary mean declines with its wedge. The identity returns in Section 6.1, where adding the school’s mean Grade 8 mark, which contains the wedge, reverses the sign of the wedge’s coefficient (Appendix M).

Figure 4 shows the 2022 wedges of all six provinces in two ways, by the school’s raw Grade 7 quintile within its province and by its national poverty quintile. It then plots the share of schools in each group whose assessments are easier than their province’s median school’s. Schools with easier assessments exist in every raw-mark quintile, so easier assessments are not equivalent to weakness. The two cuts run in opposite directions. Difficulty falls with the raw mark, per the bridge identity, and rises with wealth. Almost all of this rise is at the top: the richest quintile of schools has far harder assessments than the rest. Some of this could be intake (Section 3.1), but the gradient also survives a Grade 7 slope fitted for each poverty quintile and a later Grade 8 term (Appendix H.8); Section 4.5 takes it up.

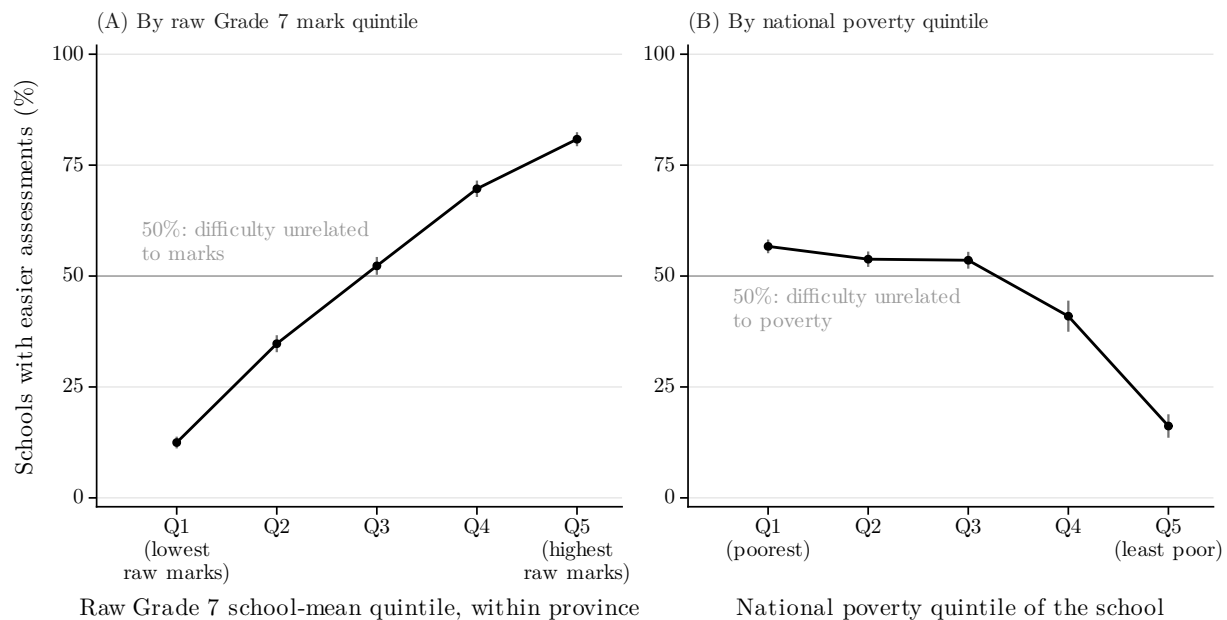


Figure 4: **Schools with easier assessments appear at every raw-mark quintile, and the richest schools have the hardest.**

Notes: Each point is the share of schools whose wedge lies below their province’s median in the 2022 cohort, that is, schools whose assessments are easier than the median school’s, by the school’s raw Grade 7 mean quintile within its province (Panel A) and by its national poverty quintile (Panel B). Bars are 95% confidence intervals for each share p , from its standard error $\sqrt{p(1-p)/n}$, where n is the number of schools in the group. The grey line at 50% is the share if difficulty were unrelated to the cut. The wedges are the province-by-province estimates of Section 3.2. Panel B leaves out the 591 schools with no national poverty quintile. **Sample:** the six provinces, mathematics, 2022 cohort: 11,890 schools in Panel A and 11,299 in Panel B.

The rest of the paper exploits the primary school’s wedge, the student’s calibrated score, and the primary school’s calibrated learning level. Section 4 ranks schools on their calibrated level. The student-level score is used both in Section 4.7, which shows that calibration improves prediction of the National Senior Certificate, and in Section 6.2, which shows how calibration permits value-added estimation.

4 Results

Section 3 put every primary school’s marks on one scale. Calibration changes the conclusions of applied work on learning outcomes. Schools differ widely in how hard their assessments are (Section 4.1), and those differences are measured precisely and persist (Section 4.2). Most schools change quintile on the common scale. The re-ranking of schools runs throughout the distribution, not just the noisy tails (Section 4.3). Replacing the raw school mean with the calibrated learning level in a simple learning production function moves its coefficients, and across the six provinces raw marks understate the poverty gradient (Section 4.4); poorer schools set easier assessments (Section 4.5). In the same provinces, raw marks show the gap between the richest and the poorest primary schools roughly halving between the 2018 and 2022 Grade 7 cohorts, while on the common scale it did not narrow (Section 4.6). Five years after primary school, the calibrated score predicts the school-leaving examination better than the raw mark does (Section 4.7).

4.1 How much schools differ

Primary schools differ widely in how hard their assessments are. The standard deviation of the wedge across primary schools is 0.47 in both 2018 and 2022. This is nearly half the within-province standard deviation of students’ Grade 8 marks: about 8 marks out of 100, compared with about 18. The gap between schools with harder and easier assessments is correspondingly large. Averaged over the six provinces and both cohorts, schools at the 10th and 90th percentiles of difficulty differ by 1.13 SD (Section 3.2). Two students with the same Grade 7 mark, one from each of these schools, would therefore score about 20 marks apart in the same secondary school. Assessment difficulty varies slightly more across schools than learning does: the standard deviation of schools’ calibrated learning levels is 0.42 in 2018 and 0.45 in 2022 (Section 3.4). The dispersion is similar in every province. Estimated separately for each of the six provinces in each cohort, the standard deviation of the wedge ranges from 0.43 to 0.49.

Dispersion is not due to noise and is not fully explained by poverty. First, two estimates of a school’s wedge, each from a random half of its students, agree closely (Section 4.2). Second, schools differ more from one another within poverty, fee-status and location categories than the categories differ from each other (Section 3.4). Third, measurement error in the Grade 7 mark, if anything, biases the estimated dispersion downward (Section 3.1).

4.2 Precision and persistence

Reliability of the calibrated school effect is measured by the Kane–Staiger signal share (Kane and Staiger 2002). Each school’s effect is estimated as an average over its students, so it comes with estimation noise, and because it is an average, that noise partly cancels, more so at larger schools. The variance in $\hat{\alpha}_j$ across schools therefore has two parts: genuine differences in school quality, and the estimation noise that remains after averaging. Reliability is the first part as a share of the total: I take the total between-school variance, subtract the average estimation error (the mean squared standard error, computed from the fit’s full covariance matrix with each student’s error variance estimated leaving that student out (Kline et al. 2020)), and divide by the total. This share of the variance is 0.95 in 2018 and 0.95 in 2022, so estimation error accounts for about 5 per cent of the differences the measure records between schools (full decomposition in [Appendix H.7](#)).

A more demanding check splits each school’s students at random in half and re-estimates the calibration on each half entirely independently. If the measure tracked which students a school happened to enrol, the two halves would often diverge. Instead, even with the intrinsic increase in noise, the two halves agree closely. The split-half correlation between the independent estimates is 0.88 in 2018 and 0.88 in 2022. A Spearman–Brown projection to the full sample lifts this agreement to about 0.94. Split-sample reproducibility is more conservative than Kane–Staiger reliability as it requires stability under a change in students, not just precision at the observed sample. Both halves share the cohort and the Grade 8 term the fit uses, so their agreement does not show that the wedge lasts. Net of estimation noise, about 66% of its variance recurs in the other cohort, about 33% is specific to the cohort though shared by all four Grade 8 terms, and about 2% is specific to the Grade 8 term ([Appendix G](#)). This is the measure the paper carries forward.

The parametric alternatives (quadratic, piecewise-linear, empirical-Bayes, and cross-classified) recover the same ordering, with each of their effects correlating with the baseline at 0.987 or above (Pearson); [Online Appendix Table 25](#) reports them beside three estimators of other objects.

The measure is also not sensitive to student demographics. Re-estimating the calibration with student over-age, gender and race controls leaves the ranking essentially intact, and wedges estimated separately on a school’s over-age and on-age students, or on its girls and boys, order schools alike, differing in spread only as much as the groups’ Grade 7 slopes and marks differ ([Appendix S.6](#)). These residualised effects correlate with the baseline at 0.999 or above across all three in both cohorts. School poverty quintile is different in kind. It is fixed within each primary school, so the school effect already absorbs it and it cannot enter the calibration as a student control. Removing each quintile’s mean from the school effects leaves a measure that correlates with the baseline at 0.92 in 2018 and 0.82 in 2022, because part of the wedge runs along the poverty gradient ([Section 4.5](#)).

Between the cohorts, 2018 and 2022, the calibrated levels correlate at 0.889 (Spearman 0.758), and the wedges on their own at 0.624. Even though there is no between-year anchor, this level of stability suggests that the estimate reveals something about the school itself. The COVID-19 pandemic fell within those four years. The calibrated levels correlate substantially more than the raw marks, which correlate at 0.441 ([Table 5](#)). A school’s 2018 Grade-7 mean predicts its 2022 calibrated standing at only 0.316. For a household or administrator trying to anticipate a school’s future standing, the calibrated level is a more stable construct than raw marks.

Table 5: A school’s calibrated learning level in 2018 predicts its 2022 level far better than its raw Grade 7 mark does. *Prospective horse race across the four-year gap between the 2018 and 2022 cohorts. Each cell is the precision-weighted correlation between a 2018 measure (row) and a 2022 measure (column), with the share of schools changing quality quintile in parentheses. The diagonal is a measure’s own persistence and the off-diagonal is cross-measure. Precision weights are the inverse summed sampling variances of the two cohort estimates; the calibrated learning level is that of [Section 3.4](#), and the wedge on its own correlates at 0.624 across the same schools ([Appendix G](#)). **Sample:** Gauteng core, 1,280 schools present in both cohorts.*

2018 measure	2022 raw Grade 7 mean	2022 calibrated learning level
Raw Grade 7 mean	0.441 (68%)	0.316 (72%)
Calibrated learning level	0.269 (76%)	0.889 (51%)

Even though persistence is improved through calibration, schools still change rank substantially between cohorts. Between the two calibrated cohorts (2018 and 2022), 50.5 per cent of schools land in a different quality quintile. That is fewer than the 72.5 per cent that calibration moves out of their raw-mark quintile within a cohort ([Section 4.3](#)), but still about half of schools. As each cohort is calibrated within its own province and year, there is no attempt to preserve the scale across years. The gap between the cross-cohort correlation and unity mixes cohort composition, drift in assessment difficulty, estimation noise, and change in the feeder-to-destination network ([Appendix G](#)). Nonetheless, a household would have a better prediction of a school’s future standing based on the calibrated result than the raw marks of their child’s prospective school.¹³

The wedge’s own persistence is well short of constancy: 39 per cent of the variance is shared between cohorts, and [Appendix P](#) returns to the reversion that accounts for the rest.

The wedge also reproduces across subjects. The English calibration was too noisy to carry as a measure ([Appendix C](#)), but a school’s English and mathematics wedges move together: disattenuated for estimation error ($n = 842$), they correlate at 0.667. The variation sits between schools within a district rather than between districts. 89.2% of the wedge’s variance is within-district (precision-weighted, Gauteng), so districts

¹³These marks are not available to parents in the South African school system, but the case holds more generally and for education system administrators.

differ little from one another while two schools in the same district differ almost as much as two drawn from anywhere in the province. District-level rubrics or district-wide assessments would show up as within-district correlation in the wedge, and none is observed.

4.3 Which schools change rank

Put on the common scale, most schools change quintile. Figure 5 sorts primary schools into quintiles twice within each cohort, once by their mean Grade 7 mark and once by their calibrated learning level. If every school graded alike, the two sorts would agree and every school would sit on the figure’s outlined diagonal. Instead, 72.5% of the 2,903 school-cohorts lie off it. No quintile on raw marks keeps even half its schools. Only 27.0% of the schools in the bottom quintile on raw marks are in the bottom quintile of learning, so support aimed at the bottom quintile on raw marks would miss nearly three in four of the schools whose students have learned least. At the top, 41.6% stay. In the middle, where schools are packed most closely and a small change in level is a large change in rank, only 21.5% stay, and the rest move up and down in about equal numbers. In all, 38.1% of school-cohorts move two quintiles or more, and the median school moves 321 places in its cohort’s ranking of about 1,450 schools, more than the roughly 290 schools a quintile holds. The moves are not an artefact of cutting at quintiles: ranked into tenths, 60.7% of school-cohorts in 2018 and 63.7% in 2022 move two tenths or more.

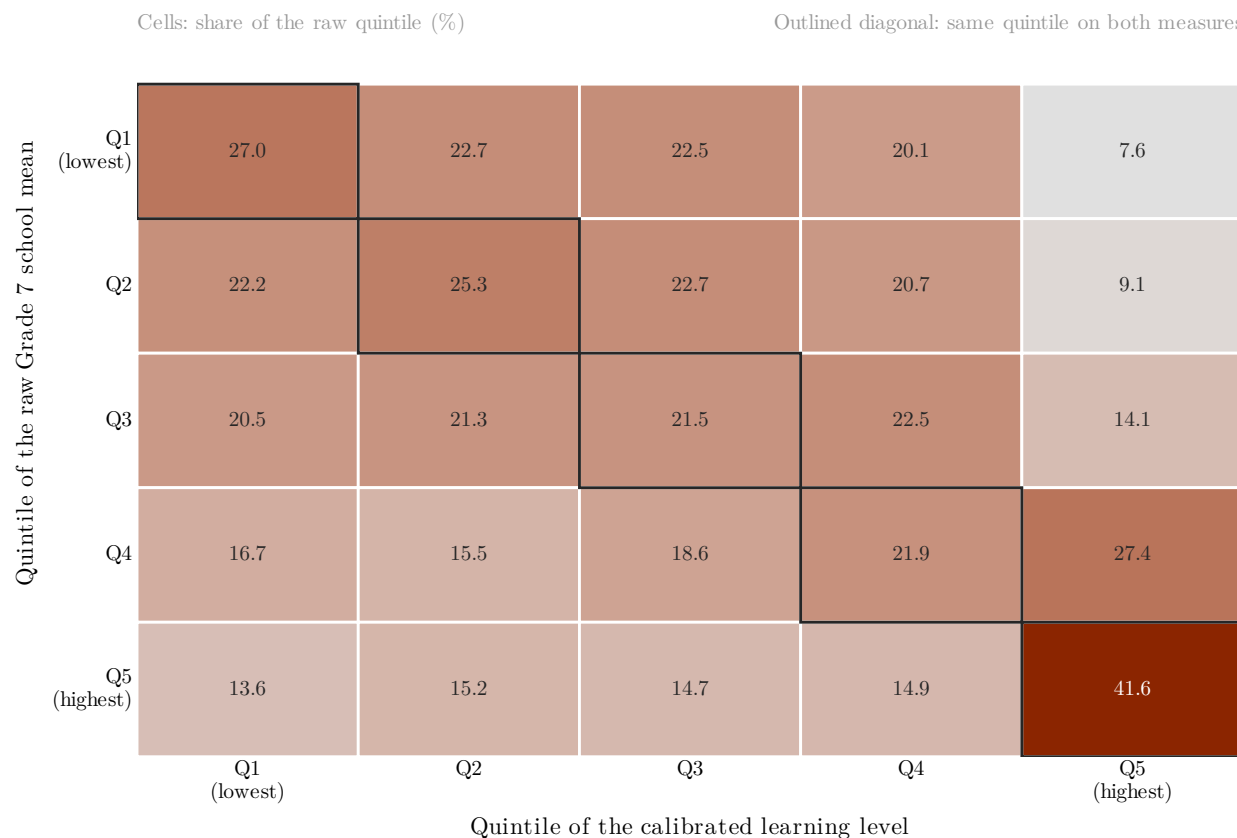


Figure 5: **Calibration moves schools out of their raw-mark quintile throughout the distribution.**

Notes: Each cell is the share (%) of the school-cohorts in a quintile of the raw mean Grade 7 mark (rows) that fall in each quintile of the calibrated learning level (columns), with quintiles formed within each cohort and the two cohorts pooled. The outlined diagonal holds the school-cohorts in the same quintile on both measures; 72.5 per cent lie off it. Rows may not sum to 100 because of rounding. Sample: Gauteng core, 2018 and 2022 cohorts, 2,903 school-cohorts.

The re-ranking is not driven by small schools. By the logic of Kane and Staiger (2002), a wedge estimated

from a few leavers is noisier than one estimated from many, so if noise drove the re-ranking, the schools with the fewest leavers would move furthest. Instead, the number of a school’s leavers explains only 0.2% of the variation in how far it moves (on a log scale), and the rank correlation between the two is -0.03 . Small schools move about as far as large ones.

4.4 What a production function recovers

Replacing the raw mean mark with the calibrated learning level as the dependent variable in a learning production function changes the estimated gradients. I regress the raw mean Grade 7 mark and the calibrated learning level separately on poverty quintile (the department’s ranking of public schools by the poverty of the community around them, from Quintile 1, the poorest, to Quintile 5), log enrolment, log class size and the student–teacher ratio, with cohort effects, for the 2,332 of the 2,903 school-cohorts for which every characteristic is recorded. Raw-mark coefficients are in standard deviations of students’ Grade 7 marks, and calibrated ones in standard deviations of their Grade 8 marks, where Section 3 measures the wedge. To compare them, I multiply each raw coefficient by persistence for its province and cohort (Section 3.2). Figure 6 reports the shift between the raw and calibrated coefficients, from the same regression with one more control: over-age, the average number of years by which a school’s Grade 7 learners are older than the age expected for the grade. The coefficients on the two richest quintiles, log enrolment, log class size, the student–teacher ratio and over-age differ significantly between raw marks and the calibrated level, the 95 per cent bootstrap interval of each difference excluding zero; those on the second and third quintiles do not.

The shifts are also large, and the largest is in the poverty gradient. The Quintile 5 advantage rises from 0.14 SD under the raw mean to 0.49 under the calibrated level, so raw marks show only 29 per cent of it, under a third. Across the provinces, the richest schools have the hardest assessments (Section 3.4). The same comparison as above, pooling the six provinces with province-by-cohort fixed effects, shows 0.24 against 0.68 SD, or 36 per cent: about a third. Evidently, raw marks understate rich schools’ students’ learning most. Over-age is excluded because assessment difficulty in part determines whether a student is held back and therefore over-age. In Figure 6, over-age’s own coefficient shrinks from -0.28 on raw marks to -0.08 on the calibrated level. Adding a control for over-age lowers the raw Quintile 5 coefficient by more than a quarter and the calibrated one by a few per cent, so raw marks then show about a quarter of the gradient: 22 per cent here and 27 across the six provinces. Class size moves too, in Gauteng and across the six provinces: its coefficient grows from -0.21 to -0.50 SD per log point, and the difference excludes zero.

Because the calibrated level is itself estimated, the wedge carries first-stage uncertainty that must be accounted for when the estimate is used in a second-stage estimator. Standard errors clustered at the primary school level would understate uncertainty. To ensure the first-stage error is accounted for, I use a paired cluster bootstrap with $B = 1,000$ replicates, resampling schools (so a school’s 2018 and 2022 observations move together) and redrawing each wedge within every replicate before re-fitting. Confidence intervals drawn from both the naive clustered standard errors and the bootstrap replicates are reported in Figure 6. These bootstrapped intervals are not substantially larger and do not change the significance of any coefficient due to the overall precision of the wedge estimates (Section 4.2). Nonetheless, work which uses the calibrated level should still be careful to account for the first-stage error.

4.5 Which schools set easier assessments

Of a school’s own circumstances, poverty is what reliably separates schools with easier assessments from schools with harder ones. A Romano–Wolf stepdown adjusts each covariate’s p-value for the number of hypotheses tested in its family. Among the school’s resources and structure, two covariates survive it: the no-fee indicator, at an adjusted p below 0.001, and the school’s own poverty quintile, at 0.008. Both say that poorer schools have easier assessments. The covariates jointly explain 37.1% of the variance in the persistent wedge (Gauteng matched, $n = 1280$), so the pattern lives in the combination of a school’s circumstances as much as in its poverty alone.

Every covariate in the atlas is standardised, so each coefficient reads as how many cross-school standard deviations easier a school’s assessments are per standard deviation of the covariate. On its own the no-fee indicator carries 0.490; with the school’s own poverty quintile, its other resources and the mechanical controls

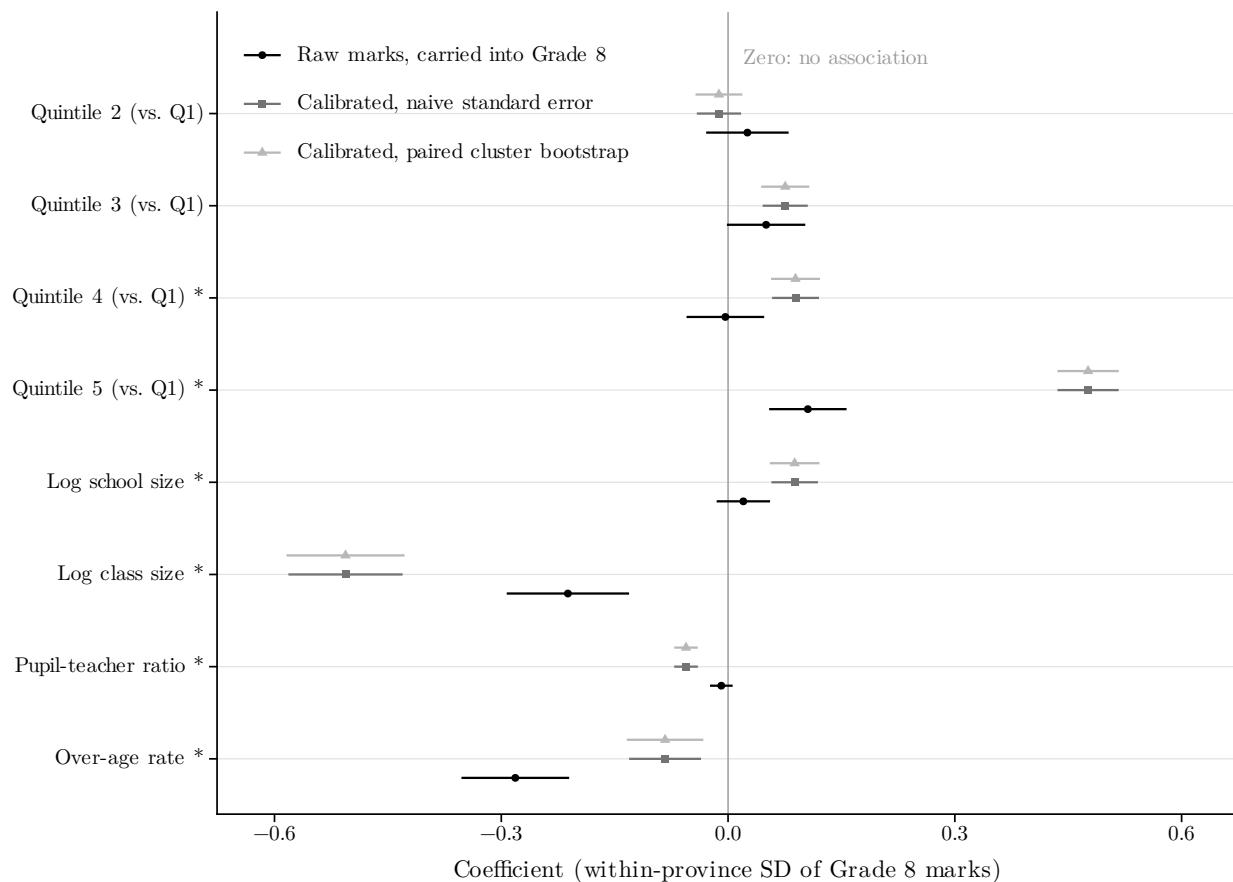


Figure 6: **Calibration moves the poverty gradient most; the bootstrap barely widens the intervals.**

Notes: Rows are coefficients from a regression of the school's raw mean Grade 7 mark, 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 text's gradient 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 marks. Bars are 90 per cent intervals; a star marks a shift whose 95 per cent bootstrap interval excludes zero. Sample: Gauteng core, 2018 and 2022 cohorts, 2,332 school-cohorts in 1,283 primary schools.

held fixed, it carries 0.426 (Figure 7). Re-estimating the wedge with poverty and achievement interactions leaves the no-fee gradient between 0.99 and 1.01 times its single-slope value. Appendix P reports the full covariate effects and the variance decomposition behind the covariate families.

Two mechanical explanations for the gradient fail. If the wedge measured differential fade over the transition rather than difficulty, the no-fee gap would change with the anchor horizon; re-estimated on Grade 8 Term 2, Term 3 and Term 4 marks, it is essentially flat (Appendix H.8). If it measured strategic inflation at the pass mark, the schools that heap at the cut-points would be the ones with easier assessments; theirs are harder on average. Where marks pile up just above a threshold, the school is on average marking below the anchored learning. That correlation has three possible readings, and the behavioural one is the least secure. One is arithmetic: a school whose marks sit low has more students just below 40, and so more marks to push over the line (Appendix B.5). Another is institutional: heaping follows the promotion rule (Section 5.2), which fixes when it happens. Only what those two leave over can be behavioural, such as heaping as a countermeasure at a school whose standard sits above the province's.

Shapley values divide the explained variance among the covariate families. The outcome is the persistent wedge, the standardised cross-cohort average of $\hat{\alpha}_j$, and the shares divide the 37.1% joint R^2 , not the total variance of the wedge. The three largest families sit close together: downstream market structure contributes 38.3%, resources and structure 28.8%, and accountability exposure 28.4% (Online Appendix Table 51 lists the variables in each). Downstream market structure leads without dominating, which matches the covariate atlas: no single family explains the pattern on its own.

4.6 The rich–poor gap across cohorts

Raw marks also distort the gap between rich and poor schools and the change in that gap over time. Across the six provinces, Figure 8 compares the mean mathematics achievement of primary schools in the richest poverty quintile with that of schools in the poorest in the same province, for every Grade 7 cohort the data can calibrate; the 2020 and 2021 cohorts are calibrated exactly as 2018 and 2022 are, province by province.¹⁴ On raw Grade 7 marks, the richest schools' lead fell from 0.48 SD in 2018 to 0.27 in 2020, 0.26 in 2021 and 0.23 in 2022.¹⁵ On the common scale, the rich–poor gap did not narrow: it was 0.72, 0.66, 0.71 and 0.77 SD in the same cohorts. The raw marks show a dramatic decline in the gap, when, in fact, the calibration shows that no such decline has occurred. TIMSS, whose scale is linked across cycles, sides with the common scale: between 2019 and 2023 the mathematics gap between the richest and the poorest quintiles of schools held steady in Grade 9 and widened in Grade 5 (Reddy, Winnaar, Harvey, et al. 2022; Reddy, Winnaar, Arends, et al. 2022; Department of Basic Education 2024).

The narrowing of the raw gap between the cohorts is plausibly due to pandemic-era policy of increased leniency. Van der Berg et al. (2024, 14–16) report that this leniency raised pass rates in 2020 across all provinces and subjects, and relieved learners most in poorer provinces and districts, especially in Grade 10. The Grade 12 leaving assessment allows the effect of this leniency to be tested. They report the Department of Basic Education's estimate that three-quarters of the 2022 rise in NSC passes above pre-pandemic predictions came from changes to the promotion rules rather than from changes in learning.

4.7 Validation against the National Senior Certificate

A new method for calibrating school-based assessment would ideally be validated against a comparable external assessment. No such assessment exists in Grade 7 or Grade 8, so I turn to the National Senior Certificate (NSC), the school-leaving examination, which I can link for one cohort: students in Grade 7 in

¹⁴The value-added estimates of Section 6.2 take their 2020 calibration from a link on every term of both years rather than on the end-of-year Grade 7 and second-term Grade 8 marks alone; on that link the 2020 gaps are 0.26 SD on raw marks and 0.67 SD on the common scale.

¹⁵Each cohort's marks are standardised within its own province and year, so each gap is measured against the spread of that cohort's students: in 2018 the average primary school in the richest poverty quintile stood 0.48 of a standard deviation of its cohort's Grade 7 marks above the average school in the poorest, and in 2022 0.23. No test links one cohort to the next, so these are gaps relative to each cohort's spread, comparable across cohorts so long as students' learning is about as widely spread in each. A change in that spread would move the raw and the calibrated gaps in the same direction; it could not make the raw gap halve while the calibrated gap holds.

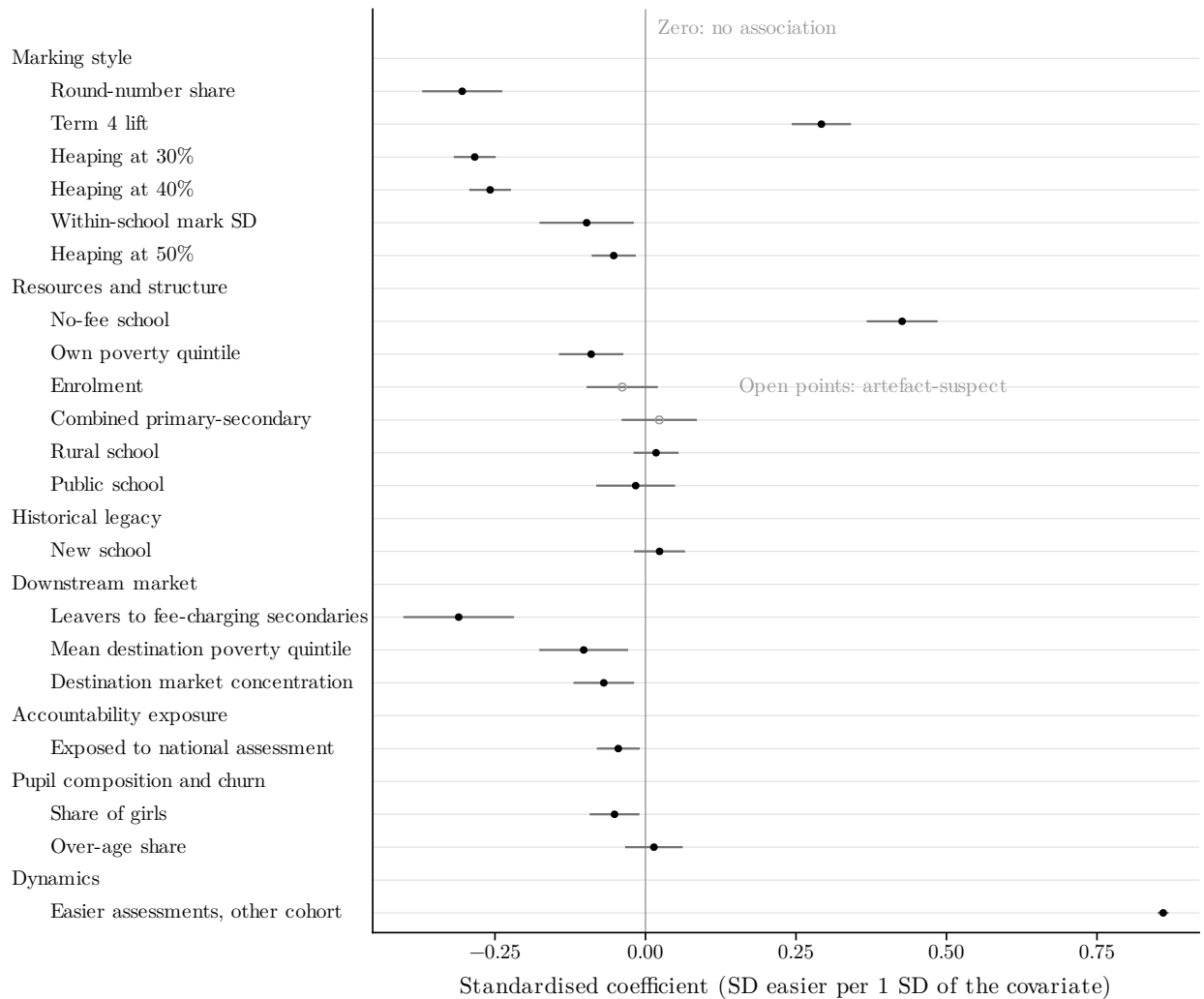


Figure 7: **No-fee status carries the largest association among school characteristics with easier assessments.** Each row is one covariate, grouped into family panels; coefficients from separate precision-weighted OLS regressions, each conditional on the mechanical and resource covariate families, the school's own poverty quintile among them, and cohort fixed effects. Every covariate and the outcome are standardised, so a coefficient is cross-school standard deviations of the outcome per standard deviation of the covariate. The outcome is the sign-reverse of the wedge, so a positive coefficient is easier assessments. Horizontal bars: 95% confidence intervals. The no-fee indicator carries 0.426 (0.490 on its own). The dynamics row, the same school's wedge in the other cohort, is larger but mechanical, and is excluded from the variance decomposition for that reason; the covariate families jointly explain 37.1% of the variance in the persistent wedge ([Appendix P](#)). **Sample:** Gauteng Mathematics, up to 2,753 pooled school-cohorts with an own poverty quintile (1,280 schools seen in both cohorts; precision-weighted). Associations only; no causal interpretation. [Appendix P](#) carries the full covariate atlas and the variance decomposition.

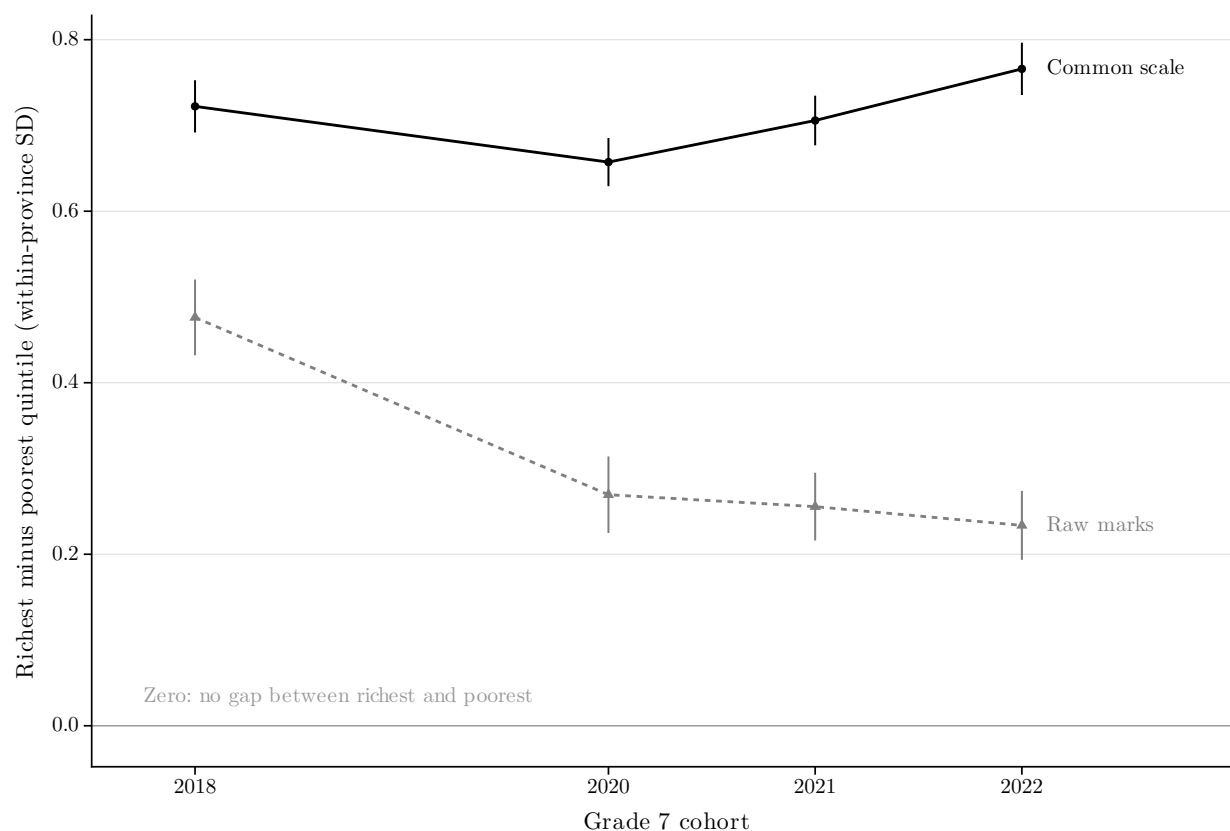


Figure 8: **The rich–poor gap roughly halved on raw marks and held on the common scale.**

Notes: Each point is the gap in mean Grade 7 mathematics achievement between primary schools in the richest and the poorest poverty quintiles of the same province, with no controls, cohort by cohort. Each cohort is calibrated on its own link from end-of-year Grade 7 marks to second-term Grade 8 marks. Raw gaps are in within-province SD of Grade 7 marks, calibrated ones of Grade 8 marks. Whiskers are 90 per cent intervals clustered by primary school, which treat the wedge as observed. Poverty quintiles are from the 2024 school masterlist. The 2019 cohort is not shown: its Grade 8 year, 2020, recorded almost no second-term marks. Sample: the six provinces pooled; 10,396 to 11,299 primary schools a cohort.

2020, to their results in 2025. Although five years separate the two, whether the raw mark or the calibrated score better predicts a student’s NSC result is a strong test of whether the calibration does real work. In each of the six provinces and across them together, in both NSC Mathematics and Mathematical Literacy,¹⁶ the calibrated score predicts better: the correlation rises by about a quarter in Mathematics, from 0.31 to 0.39, and by about three-fifths in Mathematical Literacy, from 0.25 to 0.40.

The NSC is only partly external: the final mark blends school-based assessment, set and marked by the student’s own Grade 12 school and then moderated statistically against the examination (Umalusi 2016), at 25% with an external examination at 75%. Across the 70,790 students of the 2020 Grade 7 cohort who sat NSC Mathematics in 2025, the Pearson correlation between the NSC mark and the raw Grade 7 mark is 0.31; on the calibrated score, which combines the wedge and the student’s own mark at the weights the calibration fixes, it is 0.39. The Grade 8 mark gives the mirror image. Its calibrated version is the Grade 8 mark 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 mark correlates more strongly in 14 at Grade 7 and in 14 at Grade 8. No weight in either calibrated mark was fitted to the NSC. These correlations hold nothing fixed, so part of the gain could come from the secondary school a student enters, through what it teaches or, for students who stay, through the quarter of the NSC mark it assesses itself. The tougher check, of the score as a measure, holds the secondary school fixed, and it is possible at Grade 7 only: the secondary school of entry is the Grade 8 school, so its fixed effect would absorb γ_k and leave the raw and calibrated Grade 8 marks identical. I regress each student’s NSC mark on her Grade 7 mark, with a fixed effect for the secondary school she entered and the same controls in every fit, and compare three versions of the Grade 7 mark: the raw mark, the calibrated score, and a version in which the data choose the weights freely. The calibrated score fits better than the raw mark in 14 of the same 14 cells, and the within-R-squared rises from raw to calibrated to freely weighted in 14 of them. In Mathematics it goes from 0.152 on the raw mark to 0.191 on the calibrated score, and freeing the weights adds only 0.005 more; in Mathematical Literacy the three figures are 0.160, 0.204 and 0.003. The raw and calibrated fits are not nested, so this is a comparison of fit, not a formal test. The secondary-school fixed effects of the calibrated-score fit are the exit-school effects of Section 6.2. The school-level comparison points the same way. Across secondary schools, the Spearman rank correlation between the Grade 8 intake mean and the school’s own NSC mean is 0.23 for the raw intake mean and 0.30 for the calibrated one in the 2018 cohort, and 0.26 against 0.33 in 2022, a calibration gain of +0.07 and +0.07 (Appendix K). That benchmark is noisy: it is a single year’s school mean, from the 2022 examination for both cohorts, written by different students from the intake it is compared with.

The school-level comparison above sets one aggregate against another. One cohort, in Grade 7 in 2020 and sitting the NSC in 2025, allows a student-level test of the wedge itself. Among students who entered the same secondary school with the same raw Grade 7 mark, those from an origin school with harder assessments do better five years later. They do so first on whether they sit at all. On the whole entering cohort of 394,353 students, a one-unit higher origin wedge, one within-province standard deviation of the Grade 8 mark, raises the probability of sitting the NSC by 0.340, against 0.196 for a one-unit higher own Grade 7 mark, one standard deviation of the Grade 7 mark, from a base of 0.396 (Figure 9, Panel B). The two steps differ in size. In a student’s calibrated score, a standard deviation of the wedge counts as much as 1.91 standard deviations of her Grade 7 mark. Measured in the same steps, harder assessments raise the probability of sitting by 0.91 of what the student’s own mark does. That cohort is not selected on the wedge, which is why the participation margin leads. Among those who sat, the wedge predicts the student’s own NSC mark at 0.697 national NSC standard deviations per unit in Mathematics and 0.751 in Mathematical Literacy, against 0.458 and 0.463 per unit of the student’s own Grade 7 mark, with a fixed effect for the secondary school of entry and controls for the student and the origin school (Panel A; Online Appendix Table 32 reports the specifications in full). Measured the same way, the wedge carries 0.80 and 0.85 of what the own mark carries. Adding the origin school’s own mean Grade 7 mark 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 Section 3.4, the wedge is

¹⁶Students in Grades 10 to 12 choose between Mathematics and Mathematical Literacy, the less demanding of the two; Mathematics is required for university programmes in science, technology, engineering, mathematics and commerce (van der Berg et al. 2024, 3).

roughly its students’ mean Grade 8 mark minus a fraction of their mean Grade 7 mark, 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 marks is one with easy assessments. That specification (M3) is therefore a diagnostic. The coefficients to read are those in the fits above (M1) and on the participation margins, the same reading [Appendix M](#) gives the sorting test. The origin school’s quintile, no-fee status and rural status enter every fit as controls, so all of these coefficients are already net of them.

The calibration fixes how much larger the wedge coefficient should be than the own-mark coefficient. If the NSC mark tracked calibrated learning, their ratio would equal the reciprocal of the Grade 7 slope β_1 , 1.91. The observed ratios fall short of it: 1.74 on the participation margin, the closest of the three, and 1.52 in Mathematics and 1.62 in Mathematical Literacy on the mark ([Figure 10](#)). Measurement error in the wedge accounts for most of the participation shortfall but for the smaller part of the two mark 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 [Section 4.2](#) 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 mark 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. Instrumenting the wedge with the same primary school’s wedges from the 2018 and 2022 cohorts keeps only the part that persists, and that part predicts sitting with the full weight calibrated learning gives it: its ratio is 1.95, and the 95 per cent confidence interval contains the benchmark. The persistent part does not close the mark shortfalls. In Mathematics it moves the ratio further from the benchmark, to 1.32 ([Appendix K](#)). Two explanations fit what remains, and I cannot separate them. One is selection into sitting, taken up below. The other is that the Grade 8 mark that fixes β_1 weights Grade 7 learning differently from a mark five years later. Imposing the calibration’s weights despite the shortfall costs little fit: freeing the weights on the wedge and the own mark raises the within-R-squared by only 0.005 in Mathematics. 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. The shortfall describes the three ratios plotted, not every margin: on the margin of sitting Mathematical Literacy, the ratio overshoots the benchmark.

These fits have five limitations; [Appendix K](#) discusses the first two. First, the mark regressions use only the students who sat, a fraction 0.396 of the cohort, and since the wedge predicts sitting, that sample is selected on the regressor; nothing in this panel bounds the selection. Second, the outcome is the blended mark, a quarter of which the Grade 12 school sets and marks. Third, the records that carry the outcome link to 63.7 per cent of the 2020 calibration panel, and unevenly: to 75.6 per cent of students from quintile-1 primary schools but 23.9 per cent of those from quintile-5, a ratio of 3.2. The fits therefore lean towards the poorest origins, where assessments are easiest ([Section 4.5](#)). Fourth, the regressor is the 2020 cohort’s wedge, from a separate run of the calibration, and that cohort spent both years of the transition inside the COVID-19 disruption that the main cohorts were chosen to avoid ([Section 2.4](#)). I use it because it is the only cohort I can follow student by student to the examination: the term marks that link the two record systems begin in 2021 ([Appendix G](#)). Fifth, the wedge is a generated regressor. [Section 4.4](#) and [Section 6.1.5](#) redraw it inside a paired bootstrap, but the standard errors here are only clustered on the origin school and treat $\hat{\alpha}_j$ as known. They therefore understate the uncertainty, most where mixing networks are sparse, and this fit pools all six provinces; I do not make that correction here, and [Appendix J](#) sets out how to make it. The conditional increment and the school-level rank correlations above are consistent, because they measure different objects. The test in [Appendix K](#) 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 mark. The paper’s own 2018 cohort reaches the NSC at school level only, where its wedge predicts the school’s mark with the same sign; [Appendix K](#) reports that result.

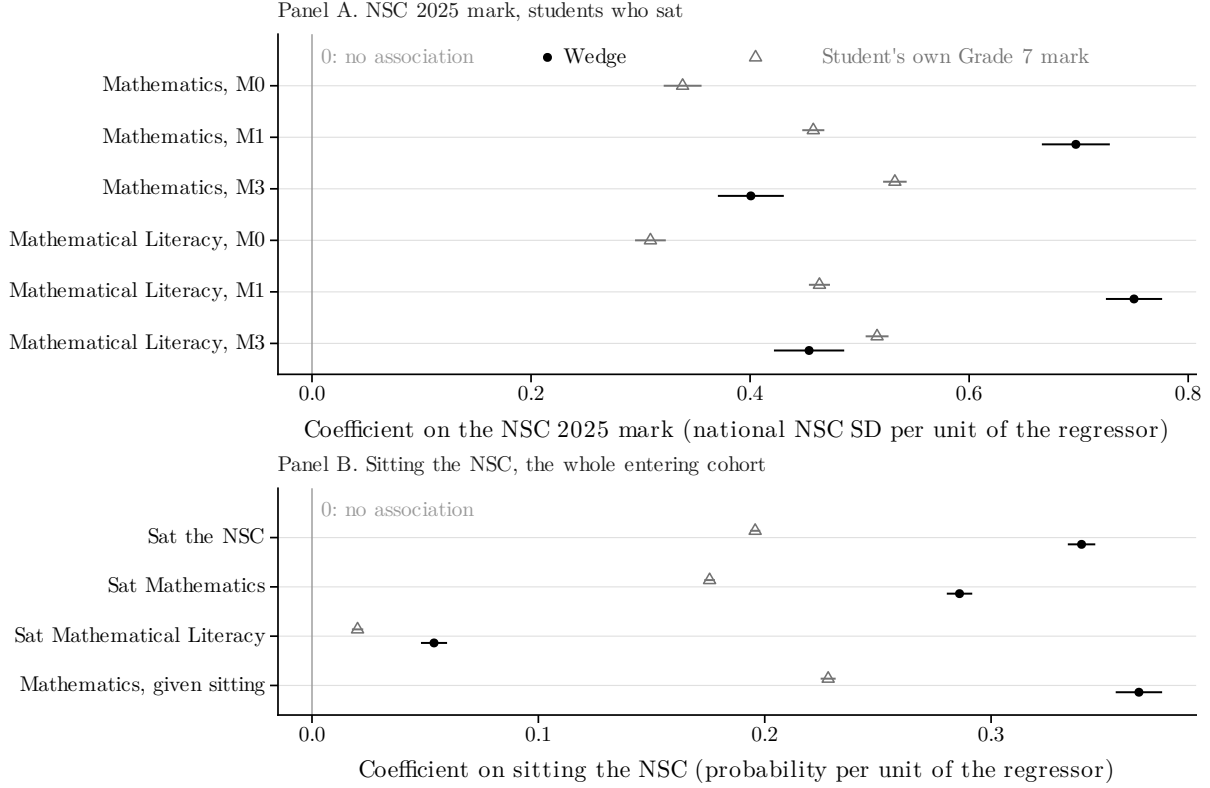


Figure 9: At the same Grade 7 mark and in the same secondary school, students from primary schools with harder assessments do better on the examination.

Note. Each row is one fitted specification, with the wedge coefficient (filled circle) and the coefficient on the student's own Grade 7 mark (open triangle) drawn side by side. The wedge is the Grade 7 origin school's estimated assessment difficulty, in within-province standard deviations of the Grade 8 mark, read from the calibration and never refitted. Panel A is the student's NSC 2025 mark, standardised nationally over every 2025 sitter of the subject; Panel B is a linear probability model on sitting the NSC, fitted on the whole entering cohort. The two panels carry different units and therefore free horizontal scales. **Coefficients are per unit of the regressor, not per standard deviation of it:** the own Grade 7 mark is standardised on the calibration sample, of which this estimation sample is a subset, so its standard deviation here is below one and reading these coefficients as per-SD effects overstates them. The wedge's coefficient is per standard deviation of the Grade 8 mark and the own mark's per standard deviation of the Grade 7 mark, so their order on the plot does not rank them. The calibration expects the wedge's coefficient to be 1.91 times the own mark's. The wedge's coefficient falls under M3, where the origin school's own mean Grade 7 mark enters as well, and the fall is built in: the origin school's mean Grade 7 mark is a component of the wedge, so M3 is a diagnostic, and the rows to read are M1 and the participation margins. Every specification absorbs a fixed effect for the secondary school of entry and controls for sex, home language, population group, over-age status, and the origin school's quintile, no-fee status and rural status. The whisker is the coefficient plus or minus 1.96 standard errors under a normal approximation; standard errors are clustered at the Grade 7 origin school. The grey vertical line at zero is the benchmark, drawn behind the data. The NSC mark is the final blended mark, 25% school-based assessment plus 75% external. 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: the two M0 rows contain no wedge by construction and are simply not drawn. Source: outputs/tables/mathematics/tab_alpha_nsc.csv. **Sample:** the 394,353 students of the Grade 7 2020 cohort who entered a secondary school in one of six provinces carrying a calibrated intake mark; Panel A is the 70,790 of them who sat NSC Mathematics in 2025 and the 85,454 who sat Mathematical Literacy, and Panel B is the whole cohort.

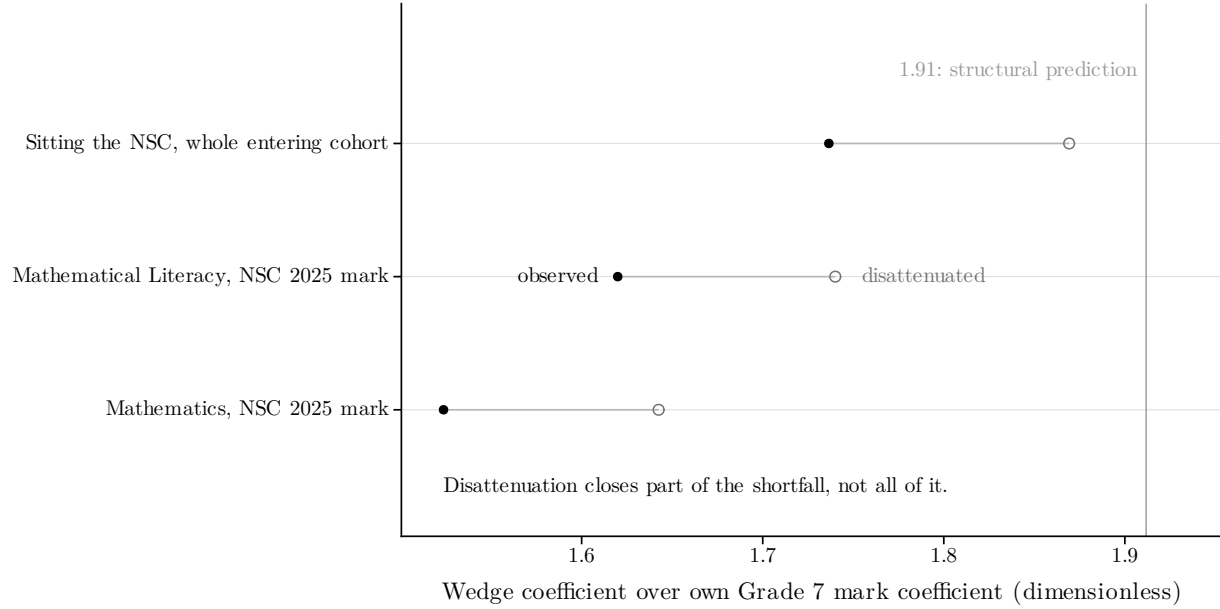


Figure 10: **Every plotted ratio falls short of the structural prediction, and disattenuation closes only part.**

Note. Each row is the ratio of the wedge coefficient to the coefficient on the student's own Grade 7 mark 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 mark, 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 mark; the two mark 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.

5 Robustness

In Monte Carlo simulations of the identification threats set out in Section 3, the ranking of schools survives where the levels do not (Section 5.1). Marks heap at the progression thresholds, most in the term that decides promotion, but removing the heaping leaves the ranking intact (Section 5.2, Section 5.3).

5.1 Simulated threats to identification

Robustness to a set of the most likely threats to identification is tested using a Monte Carlo simulation. Data are generated with known school effects where students sort on unobserved ability, where there is thin mobility between feeders and destination, where there is endogenous mobility into secondary school, and selective attrition. The ranking survives every threat, but there is degradation of the cardinal levels of the estimates. At the network densities and degree of selection observed in the data, the rank correlation with the known learning holds at or above 0.95, falling to 0.88 in only the sparsest mobility scenarios, at levels of sparsity not observed in the feeder network. The levels degrade only under sorting and endogenous mobility.

Sorting of students into classes and endogenous mobility of students selecting schools based on unobserved ability act on the spread of the estimates but leave the order or rank of the estimates intact. Both cause quality to leak between the feeder and destination effects. In the presence of sorting, the fixed effect of the feeder school absorbs some of the quality of the destination schools its students reach. Strong feeders send their students to strong secondary schools, so some of the quality that should have been attributed to the destination is instead attributed to the feeder. This spreads the estimation (“stretches the ruler”) but does not reorder the schools. Under endogenous mobility, the leak is from the secondary school fixed effect to the feeder. As the destination of the contamination is inverted, this contracts the spread (“shrinks the ruler”) but does not reorder the schools. In both cases, confidence intervals can lose coverage of the true effect, but a stretched or squeezed scale cannot place a school ahead of another (outside of estimation noise reordering). As stretching moves the true differences of schools apart while estimation noise remains constant, the chance of reversal between neighbours due to estimation noise becomes rarer. The total distortion of endogenous mobility and sorting remains small even in levels. The feeder-to-destination mixing in the data, a correlation of 0.021, would have to reach 0.048, about 2.3 times stronger, before endogenous mobility moved the ratio of the estimated to the true variance of the wedges by 5 percentage points (Appendix L.1). Ordinary estimation noise is a separate matter, and the leave-out correction of Kline et al. (2020) measures it on the data: about 5 per cent of the wedge’s variance (Appendix S.3).

Table 6: **Rankings survive all four identification threats; only cardinal levels degrade.** *Monte Carlo validation summary (rank correlations and bias). The base and severe columns report the mildest and most severe scenario for each threat. The mean-bias column reports magnitude after re-centring against known ground truth; the substantive metric is the Q5–Q1 bias spread by quintile of true τ_j . 500 replications per scenario. **Sample:** Gauteng core, 2018 and 2022 cohorts.*

Threat	Scenario	ρ (base.)	ρ (severe)	Bias (mks)	Q-stab.	Verdict
Measurement error (heaping)	Heaping index 1–3	0.968	0.968	<0.001	74.7%	Negligible
Limited mobility (sparsity)	2–11 destinations, 15–80 students	0.968	0.876	<0.001	74.6%	Negligible at observed density
Endogenous mobility (An-drews/KSS)	psi = 0 to 2, observed density	0.963	0.897	RMSE 2.51 at psi=2 (obs. density)	58.5%	Ranks stable; levels degrade
Non-random attrition	Random to extreme selection	0.960	0.960	<0.001	72.0%	Negligible

(continued)

Threat	Scenario	ρ (base.)	ρ (severe)	Bias (mks)	Q-stab.	Verdict
Sorting + peer effects	rho=0–1, lambda=0–0.15	0.953	0.985	<0.001 (RMSE 5.84 at rho=1)	79.4%	Ranks stable; levels degrade

5.2 Heaping at the promotion threshold

The Grade 7 grades that enter the calibration are themselves manipulated: marks bunch at the 30, 40 and 50 thresholds that govern student progression and achievement reporting, overwhelmingly at 40, the promotion requirement in mathematics (Section 2.2). Table 7 quantifies the excess mass at each threshold against a smooth counterfactual; at 30 and 50, which promotion does not require of mathematics, bunching is near-absent. Bunching is heaviest at the weakest schools, with the bottom fifth of schools heaping about 5.6 times as often as the top fifth (Appendix B.5). The resulting measurement error is non-classical: it is concentrated at particular grades, correlates with student attainment, and is directional, since students are pushed up to the threshold, never down. The marks bunch at thresholds, but $\hat{\alpha}_j$, the wedge between the common scale and a school’s scale, is diffuse across the grade range and is not concentrated at these thresholds (Appendix B).

Table 7: **Heaping concentrates at the 40% threshold, yet the school-level mark–learning gap is diffuse.** *Heaping severity at all three progression thresholds (Gr 7 Mathematics). The spike is at 40, the dominant threshold; marks 30 and 50 sit close to their neighbouring frequencies. Heaping index = the frequency at the mark divided by the mean of the two adjacent integer-mark frequencies. Excess mass = the departure of that frequency from a counterfactual density fitted by local polynomial to the marks that are not multiples of five. Although marks heap sharply at the 40-mark threshold, the school-level mark–learning gap is diffuse across the full mark range. **Sample:** National (6 provinces), 2018 and 2022 cohorts.*

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

Within a school grade, the Term 4 mark, which decides promotion, heaps at 40 more than the Term 1 mark of the same students in the same year. A Term 4 to Term 1 ratio of the heaping index above one means the mark was pushed to the threshold more often in Term 4, and across the plotted grades to Grade 9 the median ratio runs from 1.187 to 1.354 (Figure 11). At Grade 12, where promotion runs through the NSC rather than through the school’s own mark, the ratio is 0.988: the Term 4 excess disappears. 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. The school’s mark enters the certificate only as a quarter of the final mark, after moderation against the examination, so lifting a student across 40 on the school’s own mark buys her little. The extract records the Grade 12 mark as the school set it, before moderation, so the moderation has not smoothed the heap away. Grades 10 and 11 sit at 1.041 and 1.065; the claim rests on the Grade 12 reversal alone. Heaping follows whoever controls promotion. It is institutional rather than particular to a school or a marker, which is what the diffuse wedge described above requires.

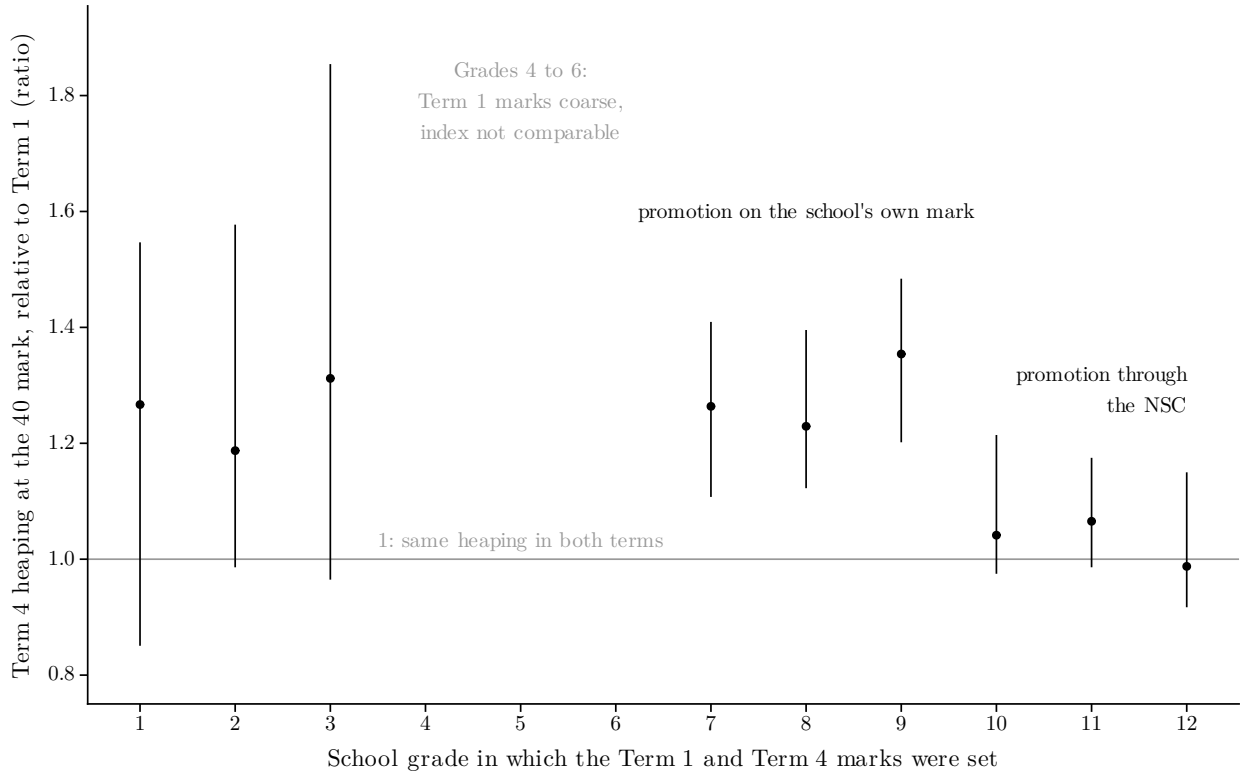


Figure 11: **Heaping tracks the term that decides promotion, and dies at Grade 12.**

Note. Each point is the median, across province-by-year cells of that grade, of the ratio between the Term 4 and the Term 1 heaping index at the 40 mark; the whiskers are the interquartile range across those cells. The heaping index is the frequency of the 40 mark over the mean frequency of 39 and 41, so a ratio above one means the mark 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 mark. **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 mark, 25% school-based assessment plus 75% external. A cell enters only where both terms use the full integer mark scale, and a grade only where at least ten such cells 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 source carries a cell 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 marks in the national school-based-assessment extract, Term 1 against Term 4 of the same year at the same school grade; 186 province-by-year cells across the 9 plotted grades, North West 2022 excluded because its two blocks duplicate each other.

5.3 Smoothing the heap

The heaping is not symmetrical: students’ marks are raised to meet a threshold and never lowered to it. Because heaping happens only at certain points of the distribution, the chance of being heaped correlates with a student’s learning level, so it does not wash out across the cohort. I attempt to recover the degree of bias through a Monte Carlo simulation. A data-generating process with injected heaping at just above the observed severity (index 2.4 at the 40 threshold) with known school effects is passed to this paper’s estimator. The estimator recovers the order as well from the heaped marks as the clean ones: the rank correlation with the truth holds at 0.968, the same 0.968 as with no heaping, and the persistence slope shifts by only 0.003 between heaped and clean marks, even under heaping pushed beyond the observed level. When the data is generated with a correlation between the degree of heaping and school quality, at or above the correlation observed in the data, the rank correlation with the underlying quality stays within 0.0004 of the clean-mark value and the persistence slope moves by under 0.4 per cent (Appendix L.5).

A de-heaping exercise on the observed data likewise recovers a materially unaltered ranking of schools. Smoothing both mark distributions with a Kleven–Waseem counterfactual density (Kleven and Waseem 2013) (Figure 12) and re-estimating moves the persistence slope β_1 by at most 2.1 per cent in any province–cohort cell, and leaves raw and de-heaped $\hat{\alpha}_j$ correlated at 0.9973 or above throughout. That counterfactual is a global polynomial, and Figure 12 shows it fitting the region below 40 poorly, so it reallocates mass across the whole distribution rather than only out of the spike. The exercise is therefore a robustness bound on the ranking, not a reconstruction of the unheaped distribution. The downstream regressions in Table 8 carry the same covariate gradients under the de-heaped measure as under the baseline.

Table 8: **The choice of measurement-error correction does not change the school-covariate gradients.** *Downstream regressions under alternative specifications. Column A: raw Grade 7 school mean, in marks. Columns B–D: the wedge alone, in Grade 8 marks, under baseline, multi-term, and de-heaped specifications respectively. Column A and columns B–D are different outcomes in different units, so compare B–D with one another, not with A. **Sample:** Gauteng core, pooled 2018 and 2022 cohorts.*

Covariate	(A) Raw	(B) Baseline	(C) Multi-term	(D) De-heaped
Quintile 2	-0.040 (0.700)	-0.444 (0.657)	-0.076 (0.641)	-0.439 (0.654)
Quintile 3	0.294 (0.634)	0.581 (0.595)	0.989 (0.581)	0.583 (0.592)
Quintile 4	-0.443 (0.621)	1.920 (0.582)	2.090 (0.568)	1.926 (0.579)
Quintile 5	0.491 (0.672)	6.432 (0.630)	6.706 (0.616)	6.454 (0.628)
Log enrolment	-0.115 (0.447)	2.359 (0.419)	2.151 (0.409)	2.338 (0.417)
Pupil-teacher ratio	-0.140 (0.043)	-0.460 (0.041)	-0.464 (0.040)	-0.460 (0.040)
Over-age rate	-25.190 (2.155)	5.943 (2.021)	6.241 (1.973)	5.743 (2.012)
R ²	0.146	0.334	0.346	0.337
RMSE	6.569	6.159	6.015	6.133

6 Applications

Each application takes the common scale as an input to a question with its own design. At the same mark, students from primary schools with easier assessments enter secondary schools with lower-achieving classmates (Section 6.1). With the calibrated score as the intake, descriptive value added becomes estimable, and two in five secondary schools change their quintile of value added when it is estimated with the calibrated intake score instead of raw marks (Section 6.2).

6.1 Sorting into secondary schools

6.1.1 The question

With no externally comparable assessment at primary level, where the calibration is identified, Section 4.7 tested it downstream, against students’ National Senior Certificate results five years after Grade 7. This

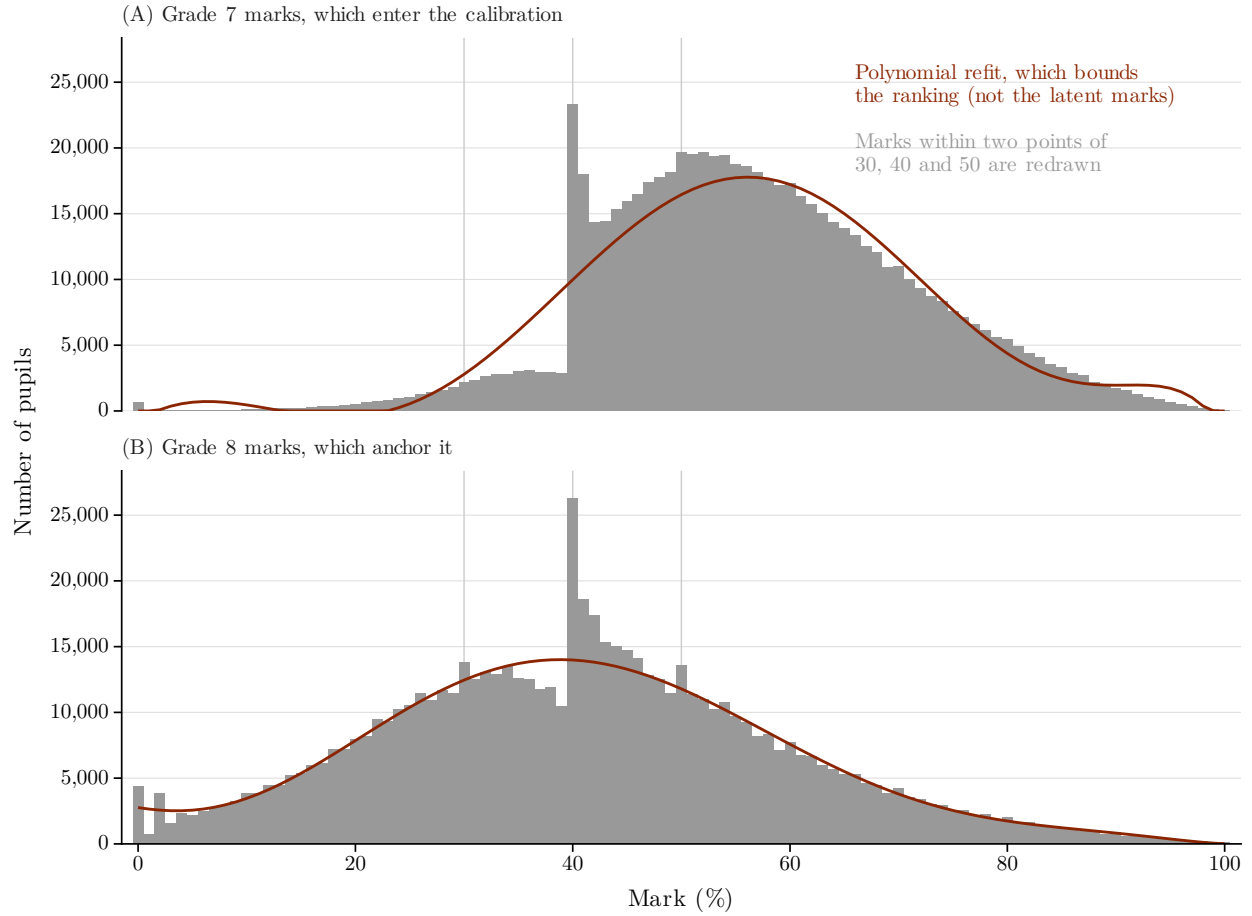


Figure 12: **De-heaping redraws the marks near each threshold from a deliberately aggressive refit.** Observed integer-mark frequencies (bars) against a polynomial refit (line), for the Grade 7 marks that enter the calibration (Panel A) and the Grade 8 marks that anchor it (Panel B). The refit is a seventh-order polynomial fit to the mark frequencies outside a two-mark window around each threshold (30, 40, 50). De-heaping redraws every mark inside those windows from the refit's shape, 25.0% of Grade 7 and 31.1% of Grade 8 marks, and leaves every other mark as observed. The refit bounds how far heaping can move the ranking of schools; it is not the unheaped distribution. Bunching is sharp at 40, the promotion requirement in mathematics, in both grades, yet the calibrated school effects move negligibly under the correction (Table 8). **Sample:** National (6 provinces), mathematics, 2022 cohort.

section tests it on a decision families make at the transition itself: where their children enrol. A wedge estimated from Grade 8 marks alone should predict nothing outside the estimator unless it captures something real about schools. The calibration of Section 3 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. 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). In the reputation models of MacLeod and Urquiola (2015) and MacLeod and Urquiola (2019) the meaning of a grade depends on the school that issues it, so the same grade buys different outcomes from different schools; the wedge is that school-by-school difference, measured directly. 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 Section 4.7, 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 F, which asked whether students from one feeder split across destinations on their Grade 7 mark and on other traits, and found that they split mainly on the mark. 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 mark 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 (Online Appendix Table 38). And it recovers the alignment whether or not any household reads a mark 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 Section 6.1.5.

6.1.2 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’ marks of Section 3: the raw mean Grade 7 mark $\bar{y}_j^{G7}$, the raw mean Grade 8 mark $\bar{y}_j^{G8}$ that the same leavers went on to earn, and the wedge $\hat{\alpha}_j$ of Section 3, positive where a school’s marks 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 (Section 3.4); 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 marks.

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

$$\bar{Y}_{k(i),-i}^z = \beta_S \mathcal{S}_{j(i)}^z + \gamma y_i^{G7,z} + X_i' \delta + \lambda_{p(k) \times c} + \varepsilon_i, \quad (5)$$

with the outcome $\bar{Y}_{k(i),-i}^z$ the leave-one-out peer mean of the cohort student i joins, measured as its raw Grade 7 intake marks or its Grade 8 Term 2 marks.¹⁷ The Grade 8 mark is written within six months

¹⁷ 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. $\mathcal{S}_{j(i)}^z$ is the one origin signal entered, $y_i^{G7,z}$ the student’s own Grade 7 mark, 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.

of secondary entry, so both outcomes measure the composition of the cohort a student joins rather than the receiving school’s later contribution.¹⁸ 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).

The test covers mathematics movers in the Gauteng core, pooling the 2018 and 2022 cohorts: 221,093 student-subject-cohort observations across 1,371 origin and 883 destination schools.¹⁹

6.1.3 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 marks her origin can predict the peer mean through those co-movers without any household choosing anything. Leave-own-feeder-out (LOFO) drops the co-movers from the peer mean and re-estimates. Table 9 applies it to all three origin signals on both destination margins, each cell a separate estimate of Equation 5.

Table 9: **Enrolment tracks the calibrated scale, not raw marks; 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 marks (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 13). The raw Grade 8 row is a mechanical benchmark: Grade 8 marks are unpublished, so no household can act on it. Controls: own Grade 7 mark, 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 core; 2018 and 2022 cohorts.*

Origin-school signal	Grade 7 intake		Grade 8 achievement	
	(1) Full	(2) LOFO	(3) Full	(4) LOFO
Raw mean Grade 7 mark	+0.310*** (0.021)	+0.109*** (0.022)	+0.023 (0.015)	+0.020 (0.015)
Raw mean Grade 8 mark	+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)

Sorting on raw marks is mostly students moving together. In the full sample the raw Grade 7 mark 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 marks into it. On the Grade 8 achievement margin (columns 3 and 4) the raw Grade 7 mark is within sampling error of zero with or without own feeders (+0.020 under LOFO). Taken at face value, a strong entry mark places a student beside like-marked 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 (Online Appendix Table 38).

¹⁸Each destination school sets and marks 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 (Section 6.1.5).

¹⁹The unit of observation is the moving student, but every signal and every outcome varies at the school level, which is where the standard errors are clustered: $\hat{\alpha}_j$ enters as an attribute of the origin school, identical for every student who left it, and the outcome is a destination cohort’s peer mean. No student-level calibrated score is formed; the only own-student regressor is the raw Grade 7 mark.

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). A student from an origin whose wedge is one standard deviation higher reaches a Grade 8 cohort whose other members, none from her school, score that many standard deviations of the peer mean higher in their first-term marks. The wedge, not raw intake selectivity, links an origin to the realised attainment of the cohort its students join. This does not yet show that the wedge measures learning. Predictor and outcome are both built from Grade 8 marks, and LOFO leaves that shared construction in place; Section 6.1.5 breaks it, Section 6.1.6 avoids it, and the student-level test of Section 4.7 never had it.

The raw Grade 8 mark, 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.

6.1.4 The negative intake coefficient: a construction artefact

The one negative cell in Table 9, the wedge’s full-sample coefficient on raw Grade 7 intake (-0.069 , $t = -2.3$), is arithmetic rather than a preference for weakly marked destinations. By the bridge identity of Section 3.4, $\hat{\alpha}_j$ is to a first order $\bar{y}_j^{G8} - \beta_1 \bar{y}_j^{G7}$, so a high-wedge school has low raw Grade 7 marks 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 marks into it. Her destination’s Grade 7 peer mean is pulled down by her own origin, and that pull is the negative coefficient.

Table 10: **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 cell of Table 9, 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 9. **Sample:** $n = 221,093$; Mathematics; Gauteng core; 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 10 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 headline cell positive ($+0.102$).

6.1.5 Threats to validity and robustness

Three threats bear on the reading, and each has a specification that answers it. Figure 13 collects the specifications; Appendix M reports the full set.

Unobserved ability. The regression conditions on the raw Grade 7 mark 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 mark 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 mark, 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 Section 6.1.4 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 Section 4.7 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 marking. A destination’s Grade 8 marks carry that school’s own grading, so a student who joins a marked-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 M 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 marks 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 cell 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 $\bar{y}_j^{G8}$, 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 (Online Appendix Figure 24; tabulated in Appendix M). Estimated province by province, the wedge’s Grade 8 LOFO coefficient is largest in Gauteng and falls by about half when all six provinces are pooled (Online Appendix Table 34), which is why the test is reported on the Gauteng core.

6.1.6 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 11). 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 Section 6.1.3 it was zero.

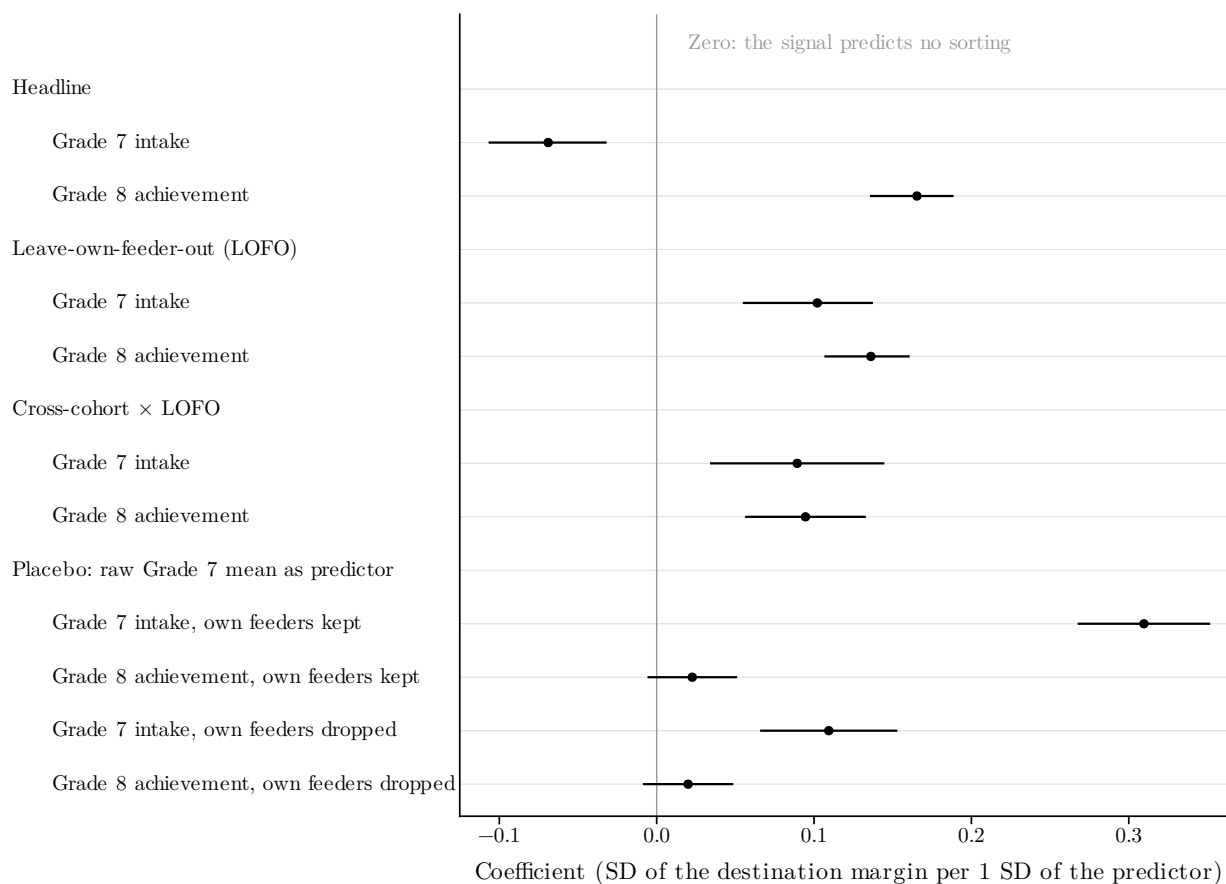


Figure 13: **Every specification keeps the wedge’s achievement coefficient; the raw-mark placebo on achievement is near zero.** Coefficient on the origin-school signal across four specifications (headline, leave-own-feeder-out (LOFO), cross-cohort $\times$ LOFO, and a placebo using the raw Grade 7 mean 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 Section 6.1.4. LOFO intervals are bootstrapped. The one point left of zero is the headline wedge coefficient on Grade 7 intake, the construction artefact of Section 6.1.4; it turns positive once own feeders are dropped. **Sample:** LOFO subsample $n = 220,973$ (99.9% of headline); Gauteng core; Mathematics; 2018 and 2022 cohorts.

Table 11: **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 9. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. **Sample:** $n = 209,666$ movers; 756 destination secondary schools; Mathematics; Gauteng core; 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 mark	+0.049** (0.015)	[0.019, 0.078]
Raw mean Grade 8 mark	+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 M](#) reports every cell. 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 mark is also the destination’s own marking: the NSC blends school-based assessment, set and marked 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 [Section 6.1.5](#) 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.

Students who leave primaries with harder assessments join Grade 8 cohorts that go on to attain more, holding fixed their own marks and the raw marks of the school they left ([Online Appendix Table 38](#)), and at the same own mark they arrive at secondary schools with stronger published matric records. In the one cohort I can follow to the school-leaving examination, they also sit it more often, and score higher on it, than students with the same Grade 7 mark who entered the same secondary school ([Section 4.7](#)). Where students enrol and how they do five years later point the same way: the wedge carries information about learning that the raw mark leaves out. A school that marks its students up therefore does not place them in a stronger cohort than their learning warrants.

6.2 Value added

6.2.1 The intake problem

A value-added measure credits a school with how far its students finish above what their intake predicts, so it is only as good as its intake. A South African secondary school’s only record of its entrants’ prior learning is the Grade 7 mark, set by each primary school on its own scale ([Section 3](#)), and on raw marks a secondary school fed by primaries with harder assessments looks as though it taught more. With the calibrated score as the intake, the secondary-school effects on the school-leaving result are descriptive value added: how far a school’s students finish above students who entered with the same Grade 7 mark. The choice of intake matters: a placebo that no secondary school could have caused, its entrants’ Grade 7 Term 1 Mathematics mark read on the intake’s own scale, disperses across secondary schools at 0.696 of the Mathematics effects’ own dispersion with the raw mark 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 mark is an external examination and the other quarter is marked by the student’s own Grade 12 school (Section 6.2.5). This setting lacks the second.

6.2.2 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) mark in Mathematics or in Mathematical Literacy, standardised nationally, so θ_k is in national standard deviations of that mark. A_i is the calibrated score of Section 3, the student’s Grade 7 Mathematics mark net of her primary school’s wedge, and X_i holds the controls of Section 4.7. 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 effects are large: net of estimation noise, their standard deviation is 0.344 (0.321), so students with the same intake at two secondary schools one standard deviation apart finish about a third of a national standard deviation apart.

The regression is the one in Section 4.7 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 mark’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 Section 6.2.5. 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 mark earned before entry as the outcome. No secondary school can have changed its entrants’ Grade 7 marks, so school “effects” on such a mark that disperse as widely as θ_k mean that the effects record who enrolls. The placebo is the Grade 7 Term 1 Mathematics mark, three terms older than the Term 4 mark 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.

6.2.3 The placebo test

The raw intake fails the placebo test and the calibrated intake passes it. With the raw Grade 7 mark as the intake, the raw Term 1 mark 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 marks still carry the primary school’s wedge.

Figure 14 holds the intake at the calibrated score and varies only the placebo’s scale. Read raw, the same Term 1 mark 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 mark tests an intake only when the two share a scale.

6.2.4 Why the raw intake fails

A raw intake credits a secondary school with its feeders’ assessment difficulty: a student from a primary with harder assessments arrives with a mark below her learning, finishes above what that mark predicts, and

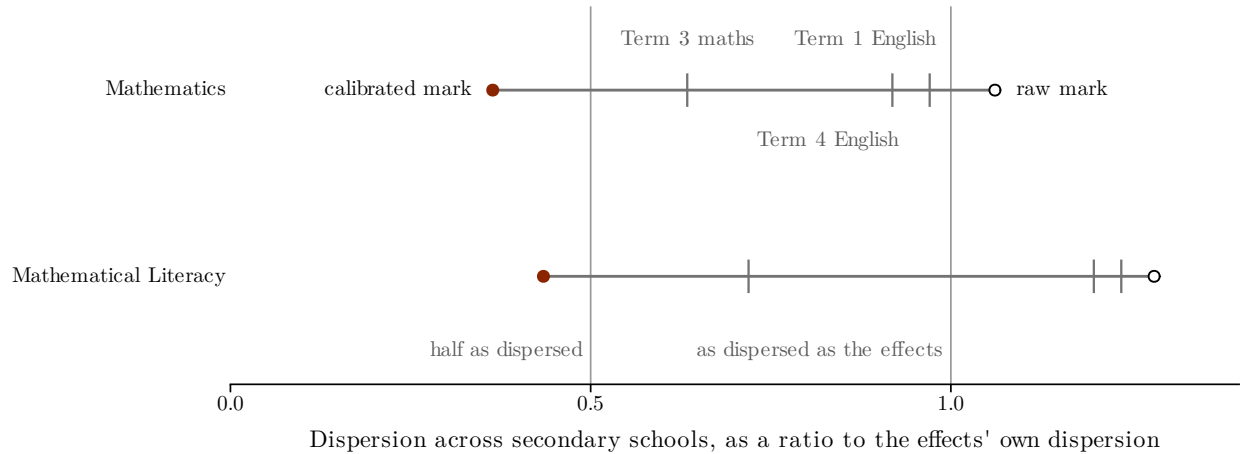


Figure 14: **Against a calibrated intake, the pre-entry mark passes the placebo test calibrated and fails it raw.**

Notes: Each marker is the dispersion across secondary schools of the school effects on a Grade 7 mark 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 mark as the intake are reported in the text and are not drawn. The open circle is the raw Grade 7 Term 1 Mathematics mark as the placebo and the filled circle the calibrated score from the same mark; 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 mark and from the Term 4 and Term 1 English First Additional Language marks. 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 mark, the final blended mark: 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.

her secondary school collects the difference. 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. 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 mark 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 mark 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 marks, 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 K](#) shows the same problem in a national school-based-assessment extract, where a single school fixed effect, net of an intake mark 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 marking out of θ_k , and leaves its early teaching in.

6.2.5 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. Specification choices that leave the intake unchanged, such as keeping only the students who sat the examination at the secondary school they entered, or only the most certain links, keep the disattenuated rank correlation with the effects at or above 0.972 (0.973); the median attrition scenario moves them most, to 0.861 (0.801). [Online Appendix Table 55](#) in [Appendix R](#) lists every check with its threshold and result.

Reliability measures persistence, and a school’s own marking persists too. A quarter of each student’s NSC mark is school-based assessment set and marked by her Grade 12 school, which for a student who stays is the secondary school she entered. Umalusi moderates that quarter against the examination school by school before it counts, pinning the school’s mean to a narrow band just above its students’ examination mean (Umalusi 2016), so a school that marks its Grade 12 students generously can raise its own θ_k only within that band, and split-half reliability can count only that much of the generosity as signal. The extract of [Appendix K](#) 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 marking, and I have not yet made that comparison. Until then, the reliability shows that θ_k is a persistent trait of the school, and not yet that it is learning.

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 ([Appendix I](#)). 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 mark lets through.

6.2.6 What the calibrated intake leaves behind

Marks 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 marks. 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.²⁰

6.2.7 What would make the effects value added

The tests that would validate the estimates as value-added cannot run on this cohort. 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 marking can run now. Until these tests run, θ_k is an exit-school effect conditional on calibrated intake, whose distribution can be compared and on which no school can be placed: 40.2% (38.6%) of secondary schools cross between above, indistinguishable and below in at least one bootstrap draw in five. The calibrated scale makes the estimation feasible, and the evidence is a proof of concept.

7 Conclusion

In most countries, schools set and grade their own assessments, so the same mark represents different amounts of learning at different schools. Often, the only comparable result is the secondary school leaving assessment. This paper derives a common scale of learning from the marks schools already award, so that a mark awarded at one school can be read against a mark awarded at another. The design requires a large number of school transitions, typical of the primary to secondary transition where demand for primary school comparability might be particularly high. Wherever students' records follow them from primary to secondary school, a system without a common primary examination can compare its schools' learning from the marks it already collects, see which schools and which groups lag and by how much, and track whether those gaps close. The Grade 7 marks the education departments already hold, once calibrated, describe a different school system from the one the raw marks show: nearly three-quarters of Gauteng's primary schools sit in a different

²⁰Both 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.

quintile of learning, and the gap between the richest and the poorest primary schools in the six provinces did not narrow between the 2018 and 2022 cohorts, though the raw marks say it roughly halved.

The difficulty wedge rescales a school's marks onto the common scale, and it carries persistent information about the school in its own right. It is reproducible in split samples, and the ranking it creates is stable, even where the level is sensitive to several possible econometric threats. Marks are pushed over the promotion threshold when the school decides promotion, not when an examination does (Section 5.2), and the pushing does not drive the wedge. The market for secondary schools corroborates the calibration. Student enrolment tracks the calibrated scale, not the raw mark, all the way up to a secondary school's leaving assessment, while the raw mark, on its own, predicts essentially none of that margin. The wedge itself reproduces from cohort to cohort. The richest schools have the hardest assessments, and the wedge is patterned by the destination markets a school's leavers enter.

The wedge also predicts what happens five years later. A student from a primary with harder assessments is more likely to sit the school-leaving examination than a student with the same Grade 7 mark who entered the same secondary school, and scores higher on it when she does; and the calibrated score predicts results in that examination better than the raw mark does. The first of these associations is the one with most at stake. A system deciding where to send support would look to students from primaries with easier assessments, who are less likely to sit, and whom the raw mark cannot tell from their classmates; the association is not causal, so it says where support should go, not what form it should take.

With linked school-leaving results, the calibrated score can serve as the intake for a value-added measure of secondary schools, and Section 6.2 shows that the estimation is feasible, though not yet that its estimates are value added. The calibrated measure itself is a composite of selection, instruction, and peer effects and not a causal estimate of learning value added. It supports comparison of schools, not accountability or any attribution of effectiveness to a school. The intake has to be one the secondary school did not set: a single school fixed effect on a school-leaving outcome, net of an intake mark the school set itself, also absorbs how hard the school's assessments were at the intake grade. In a national extract of school-based-assessment records that links secondary schools' own marks to the school-leaving result, a single school fixed effect, fitted on students who stay in the same secondary school and net of the mark that school set at intake, recovers the sum of the intake-school effect and the exit-school effect, and recovers neither of them separately (Appendix K).²¹ The intake-school effect is how hard the school's own assessments were at the intake grade, the secondary-school analogue of the wedge: the same construct, with the same sign, measured on a secondary school's own marks. The exit-school effect is how far a school's students score above or below what their intake mark predicts in the school-leaving result, and separating it from the intake-school effect takes students who change school between the intake grade and the examination. The exit-school effect so separated is not offered as an ordering of schools; estimated on random halves of each school's students, it reproduces far less closely than the wedge does. The method travels to any education system whose transition between phases mixes students from many origin schools into shared destinations. European lower-secondary transitions that track students into school types or admit them on teacher-assigned grades already hold the records it needs, and so do low- and middle-income systems where school-set marks are the only measure of learning at scale. In such a system, the records the transition creates in the ordinary course of administration are sufficient to derive a common scale of learning and each school's wedge against it. The design could also be used at the transition from secondary school to university, where, as far as I can find, no one has applied it without a common assessment. Reclassification events may offer natural experiments on what produces the gap between a school and the common scale of learning: schools changing fee status, poverty quintile,

²¹ Across 45 cells, each a grade, a subject and either a province or all provinces pooled, the Pearson correlation between the single fixed effect and the sum of the intake-school effect and the exit-school effect, taken across the schools common to both fits, runs from 0.9944 to 0.9999, close to a perfect correlation in every cell. The same correlation between the single fixed effect and the exit-school effect alone runs from 0.015, no association, to 0.607, a moderate association. The sample is students with a school-set Term 4 2021 mark in Grade 8, 9 or 10 and a same-subject school-leaving result at their own on-time examination, in Mathematics and English First Additional Language; Gauteng and the Free State carry no such mark in the extract. The split-half reliability of the exit-school effect, Spearman-Brown corrected, is 0.377 across 1,320 schools, measured for Mathematics at the Grade 8 intake on Term 4 2021 marks against the school-leaving examination of 2025. It refers to the exit-school effect as the two-way fit identifies it, from students who change school between the intake grade and the examination, and not to a single school fixed effect on the same outcome. The same statistic for the wedge in the Gauteng-core Grade 7 2018 Mathematics cohort is 0.939.

and the entrance of secondary schools in a primary school’s feeder zone.

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>.
- Andrews, Martyn J., Leonard Gill, Thorsten Schank, and Richard Upward. 2008. “High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (3): 673–97. <https://doi.org/10.1111/j.1467-985X.2007.00533.x>.
- 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>.
- Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. “Teacher Quality and Learning Outcomes in Kindergarten.” *The Quarterly Journal of Economics* 131 (3): 1415–53. <https://doi.org/10.1093/qje/qjw016>.
- Barker, Samuel E. 2026. “Measuring Human Capital Using Transcript Grades: A Factor Model in Network Data.” PhD dissertation, University of North Carolina at Chapel Hill. <https://doi.org/10.17615/96vskw06>.
- Bau, Natalie, and Jishnu Das. 2020. “Teacher Value Added in a Low-Income Country.” *American Economic Journal: Economic Policy* 12 (1): 62–96. <https://doi.org/10.1257/pol.20170243>.
- 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>.
- Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann. 2024. “Testing.” *Journal of Human Resources* 59 (2): 349–88. <https://doi.org/10.3368/jhr.0520-10886R1>.
- Bergé, Laurent R., Kyle Butts, and Grant McDermott. 2026. “Fixest: A Fast and Feature-Rich Framework for Econometric Estimations in R.” *arXiv Preprint arXiv:2601.21749*.
- 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>.
- Bloom, Benjamin S., and Frank R. Peters. 1961. *The Use of Academic Prediction Scales for Counseling and*

Selecting College Entrants. Free Press of Glencoe.

- Bold, Tessa, Deon Filmer, Gayle Martin, et al. 2017. "Enrollment Without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." *Journal of Economic Perspectives* 31 (4): 185–204. <https://doi.org/10.1257/jep.31.4.185>.
- 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, 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>.
- Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff. 2019. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." *American Economic Journal: Applied Economics* 11 (3): 382–423. <https://doi.org/10.1257/app.20170520>.
- 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>.
- Department of Basic Education. 2011a. *Curriculum and Assessment Policy Statement Grades 7–9: Mathematics*. Department of Basic Education.
- Department of Basic Education. 2011b. *National Policy Pertaining to the Programme and Promotion Requirements of the National Curriculum Statement Grades R–12*. Government Gazette No. 34600, 12 September 2011.
- Department of Basic Education. 2014. *Amended National Norms and Standards for School Funding*. Government Gazette No. 37230, 17 January 2014.
- Department of Basic Education. 2016. *2018 School Calendar for Public Schools in South Africa*. Government Gazette No. 40457, 30 November 2016. <https://www.education.gov.za/Portals/0/Documents/Publications/2018%20School%20Calendar.pdf>.
- Department of Basic Education. 2017. *School Calendar for 2019*. Government Gazette No. 41219, 1 November 2017.
- Department of Basic Education. 2020a. *Action Plan to 2024: Towards the Realisation of Schooling 2030*. Department of Basic Education.
- Department of Basic Education. 2020b. *Circular S13 of 2020: Release of the Curriculum Recovery Annual Teaching Plans for 2021*. 11 December 2020.
- Department of Basic Education. 2021. *2022 School Calendar*. Department of Basic Education. <https://www.education.gov.za/Portals/0/Documents/Publications/2022%20School%20Calendar%20Gazette.pdf>.
- Department of Basic Education. 2022a. *2023 School Calendar for Public Schools*. Government Notice 1911, Government Gazette No. 46061, 18 March 2022.

- Department of Basic Education. 2022b. *Circular S33 of 2022: Release of 2023/24 Annual Teaching Plans, Including Revised Weightings to School Based Assessment and Examinations (Grades R–12), Effective from the 2023 Academic Year*. 29 December 2022.
- Department of Basic Education. 2024. *South African 2023 Trends in International Mathematics and Science Study (TIMSS) Highlights Report*. Department of Basic Education.
- Department of Basic Education. 2025. *2024 National Senior Certificate (NSC) School Subject Report*. Department of Basic Education.
- Department of Education. 2006. *Amended National Norms and Standards for School Funding*. Government Notice 869, Government Gazette No. 29179, 31 August 2006.
- Di Addario, Sabrina, Patrick Kline, Raffaele Saggio, and Mikkel Sølvsten. 2023. “It Ain’t Where You’re from, It’s Where You’re at: Hiring Origins, Firm Heterogeneity, and Wages.” *Journal of Econometrics* 233 (2): 340–74. <https://doi.org/10.1016/j.jeconom.2021.12.017>.
- Diamond, Rebecca, and Petra Persson. 2016. *The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests*. NBER Working Paper No. 22207. National Bureau of Economic Research. <https://doi.org/10.3386/w22207>.
- Eastern Cape Department of Education. 2022. *Assessment Instruction 52 of 2022: The Weightings of School-Based Assessment and End-of-Year Examination for Grades R–11 for the Academic Year 2022*. 28 November 2022.
- Figlio, David N., and Maurice E. Lucas. 2004. “Do High Grading Standards Affect Student Performance?” *Journal of Public Economics* 88 (9–10): 1815–34. [https://doi.org/10.1016/S0047-2727\(03\)00039-2](https://doi.org/10.1016/S0047-2727(03)00039-2).
- Gustafsson, Martin. 2019. *Costs and Benefits of Different Approaches to Measuring the Learning Proficiency of Students (SDG Indicator 4.1.1)*. UIS Information Paper No. 53. UNESCO Institute for Statistics. [https://www.uis.unesco.org/sites/default/files/medias/fichiers/2025/08/GAML6-REF-7-Cost-and-benefit-approaches-to-4.1.1%20\(1\).pdf](https://www.uis.unesco.org/sites/default/files/medias/fichiers/2025/08/GAML6-REF-7-Cost-and-benefit-approaches-to-4.1.1%20(1).pdf).
- 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>.
- Himmler, Oliver, and Robert Schwager. 2013. “Double Standards in Educational Standards – Do Schools with a Disadvantaged Student Body Grade More Leniently?” *German Economic Review* 14 (2): 166–89. <https://doi.org/10.1111/j.1468-0475.2012.00563.x>.
- Kane, Thomas J., and Douglas O. Staiger. 2002. “The Promise and Pitfalls of Using Imprecise School Accountability Measures.” *Journal of Economic Perspectives* 16 (4): 91–114. <https://doi.org/10.1257/089533002320950993>.
- Kleven, Henrik J., and Mazhar Waseem. 2013. “Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan.” *Quarterly Journal of Economics* 128 (2): 669–723. <https://doi.org/10.1093/qje/qjt004>.
- 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>.
- Lance, Charles E., Marcus M. Butts, and Lawrence C. Michels. 2006. “The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say?” *Organizational Research Methods* 9 (2): 202–20. <https://doi.org/10.1177/1094428105284919>.
- 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.
- MacLeod, W. Bentley, and Miguel Urquiola. 2015. “Reputation and School Competition.” *American Economic Review* 105 (11): 3471–88. <https://doi.org/10.1257/aer.20130332>.
- MacLeod, W. Bentley, and Miguel Urquiola. 2019. “Is Education Consumption or Investment? Implications for School Competition.” *Annual Review of Economics* 11: 563–89. <https://doi.org/10.1146/annurev-economics-080218-030402>.
- Mansfield, Richard K. 2015. “Teacher Quality and Student Inequality.” *Journal of Labor Economics* 33 (3): 751–88. <https://doi.org/10.1086/679683>.
- 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.jdev-eco.2013.03.003>.
- North West Department of Education. 2024. *Annual Performance Plan 2024/25*. North West Department of Education.
- Nunnally, Jum C. 1978. *Psychometric Theory*. 2nd ed. McGraw-Hill.
- 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>.
- Reddy, Vijay, Lolita Winnaar, Fabian Arends, et al. 2022. *The South African TIMSS 2019 Grade 9 Results: Building Achievement and Bridging Achievement Gaps*. HSRC Press.
- Reddy, Vijay, Lolita Winnaar, Jaqueline Harvey, et al. 2022. *The South African TIMSS 2019 Grade 5 Results: Building Achievement and Bridging Achievement Gaps*. HSRC Press.
- Rossiter, Jack, and Maimouna Konate. 2022. *Studying School Exams: A New Database*. CGD Note. Center for Global Development. <https://www.cgdev.org/publication/studying-school-exams-new-database>.
- 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>.
- SADTU, NAPTOSA, SAOU, NATU and PEU. 2015. *ANA Adds No Value to the South African Learner*. Joint statement, 21 September 2015. <https://www.polity.org.za/article/sadtu-naptosa-saou-natu-and->

peu-ana-adds-no-value-to-the-south-african-learner-2015-09-21.

- Schwerdt, Guido, and Ludger Woessmann. 2017. "The Information Value of Central School Exams." *Economics of Education Review* 56: 65–79. <https://doi.org/10.1016/j.econedurev.2016.11.005>.
- Singh, Abhijeet. 2015. "Private School Effects in Urban and Rural India: Panel Estimates at Primary and Secondary School Ages." *Journal of Development Economics* 113: 16–32. <https://doi.org/10.1016/j.jdev eco.2014.10.004>.
- Tucker, Ledyard R. 1963. *Formal Models for a Central Prediction System*. Psychometric Monograph 10. Psychometric Corporation. <https://www.psychometricsociety.org/sites/main/files/file-attachments/mn10.pdf>.
- Umalusi. 2016. *Requirements and Specifications for Standardisation, Statistical Moderation and Resulting*. Version 5. Umalusi, Council for Quality Assurance in General and Further Education and Training. <https://www.umalusi.org.za/docs/reports/2016/REQUIREMENTS%20AND%20SPECIFICATIONS%20FOR%20THE%20STANDARDISATION%20STATISTICAL%20MODE.pdf>.
- Van der Berg, Servaas, and Debra Shepherd. 2015. "Continuous Assessment and Matriculation Examination Marks: An Empirical Examination." *South African Journal of Childhood Education* 5 (2): 78–94.
- van der Berg, Servaas, Chris van Wyk, and Martin Gustafsson. 2024. *What SA-SAMS and LURITS Data Tells Us about Education: New Insights from Administrative Data*. RESEP Report. Research on Socio-Economic Policy (RESEP), Stellenbosch University.
- Western Cape Provincial Parliament. 2022. *Minutes of Proceedings, Standing Committee on Education, 16 August 2022: Briefing by the Western Cape Education Department on How It Utilises Data*. WCPP 11/4/10.
- 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>.
- World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. World Bank. <https://doi.org/10.1596/978-1-4648-1096-1>.
- Yang, Yi. 2026. "Measuring School-Wide Grading Heterogeneity Without a Universal External Test: A Student-Course-Teacher Network with Advanced Placement Validation." *EdArXiv Preprint*, ahead of print. https://doi.org/10.35542/osf.io/9wqnv_v1.