Making School-Set Assessments Comparable: How School Assessments Hide Learning Inequality

Job Market Paper

Peter Courtney*

This version: 13 November 2026

Latest version: courtney.ac/comparable-marks

Abstract

Assessment has been found to substantially improve learning when results can be compared between schools, yet in at most a quarter of countries are all primary school students tested comparably. I recover comparable learning outcomes, without a new test, from assessments that are incomparable because their difficulty varies between schools. Because students sit their secondary school's assessments soon after arriving, I can compare primary schools through students who share a secondary school, and secondary schools through students who share a primary school. Two-way fixed effects chain these comparisons from 1.2 million South African students into a scale of grades comparable between schools. I show that between 2018 and 2022, the gap between rich and poor schools halved on school-set grades but did not narrow on comparable grades. Re-ranked on comparable grades, about three-quarters of primary schools change learning quintile. Comparable grades explain up to 2.5 times more variation in national secondary leaving examination results than school-set grades do. Students sort into secondary schools on comparable grades, not school-set grades. Estimating the value added of secondary schools requires comparable grades on entry, which my method provides.

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

*Stellenbosch University, Tinbergen Institute, and Vrije Universiteit Amsterdam. Email: p.sr.courtney@gmail.com. I thank Debra Shepherd, Servaas van der Berg, Menno Pradhan, Nic Spaull, and Magdalena Wasilewska for their guidance and comments. I also thank seminar participants at the University of Oxford, NHH Norwegian School of Economics, and Stellenbosch University for helpful comments. I thank Clare Leaver for hosting me at the Blavatnik School, where this work began. All errors are my own. The online appendix is at https://courtney.ac/comparable-marks/.

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). 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).[1] 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 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). In most countries, South Africa among them, schools set and grade their own assessments, so the same grade 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 provides a way to do exactly that.

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

The primary school's effect is its difficulty wedge, the gap between the grades it awards and its students' actual learning on the common scale. It is measured in within-province standard deviations (SD) of students' grades: positive where a school's assessments are harder than average and negative where they are easier. An assessment is easier if a given grade takes less learning to earn on it. Combining each student's grade with their primary school's wedge

[1] Their 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. The uses the World Bank review lists, 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" (World Bank 2018), need only measures that are comparable across schools, groups and subjects; none needs an assessment that is set and published nationally.

gives their calibrated score, their place on the common scale.[2] Averaging over a school's students gives its calibrated level of learning.

To illustrate: two students at different primary schools receive very different grades. Soon after arriving at the same secondary school, they sit an identical set of assessments and receive the same results, revealing, in expectation, the same level of latent learning. To have earned a higher grade in primary school with identical latent learning, the student must have faced easier assessments there. The comparison also works the other way: comparing two students who leave the same primary school with the same grade, earned on the same assessments, and enter different secondary schools shows which of those secondary schools sets and grades assessments at a higher level. 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 within-learner comparisons put almost every primary school in each province on a common scale.

Interpreting the primary school's effect as a wedge assumes that students who leave primary school knowing the same amount learn as much at a given secondary school, whichever primary school they left, and that students from the same primary school with the same grade do not sort into secondary schools by how fast they are improving. Section 3.3 tests both: wedges from the last term of the first secondary year agree closely with those from its first term (0.902), 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.[3]

The data cover the universe of students in the six of South Africa's nine provinces whose records are available. They come from the South African School Administration and Management System (SA-SAMS), in which schools record their students' grades, and each student keeps one ID from primary to secondary school. I use the mathematics grades of the cohorts that left primary school in 2018 and in 2022. Students disperse widely across schools during this transition, with the average secondary school receiving students from 15 different primary schools, which allows each secondary school's assessments to compare many primary schools at the same time. TWFE models are identified only where movers connect schools, and few movers inflate the estimated variance of the school effects (Andrews et al. 2008; Kline et al. 2020). Here, almost all students move at once, and the moves connect nearly every school in each province (98.8% or more).

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 and by cohort, so it compares schools in the same year. Results are reported across the six provinces, with each province's in Appendix E.3, and a result for Gauteng alone is named as Gauteng's.

Schools' wedges differ widely, are measured precisely, and persist. Schools at the 10th and

[2]The 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 C).

[3]The 0.902 comes from a separate national extract of school-based-assessment grades from the Department of Basic Education, Terms 1 and 4 of 2021 to 2024 in all nine provinces, not from the six provinces' linked records; it pools the extract's 2021 and 2023 Grade 7 cohorts (Appendix G.1).

90th percentiles of difficulty grade the identical underlying learning more than a full SD apart. Splitting schools' students into random halves and running the estimator on each leads to wedge estimates that correlate at 0.74 in 2022, averaged over the six provinces (0.88 in Gauteng, where the network is densest). Run the estimator on the two cohorts, four years apart, with new students, new tests, and nothing linking the years, and the wedge still correlates at 0.48 (0.62 in Gauteng). The wedge belongs to a school, not to its students or to a cohort.

As an external validation, I show that calibrated scores predict the national school-leaving examination better than raw grades do. With no common test in primary school, I turn to the first examination all students sit—the National Senior Certificate (NSC), five years later. I find that the calibrated score explains about one and a half times as much of the variation in NSC Mathematics results as the raw grade does, and about two and a half times as much in Mathematical Literacy.

The calibration changes conclusions about socioeconomic achievement gaps. South Africa ranks its public schools into five poverty quintiles by the poverty of the communities they serve. On raw grades, 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.

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 grades, because the richest schools set the hardest assessments. Ranked once by their students' mean raw grade and again by their calibrated level of learning, 77.1% of schools land in a different quintile of learning. A production function relating learning to school characteristics recovers about a third of the poverty gradient when it is estimated on raw grades. The assessments of no-fee schools, which the state bars from charging fees and which serve poorer communities, are 0.41 SD easier than those of fee-charging schools across the six provinces; at Grade 12, school grades overshoot the school-leaving examination least in the richest schools (Van der Berg and Shepherd 2015).

Easier assessment in primary school does not buy students a place at a higher-performing secondary school. At a given grade, students from schools with easier assessment enter secondary school with lower-achieving peers. A student from a primary school with assessments one SD harder joins secondary school peers with a mean grade 0.064 SD higher ($p < 0.001$), holding their own grade fixed and leaving out the peers who came from their own primary school. The primary school's raw average grade, the grade stakeholders observe, predicts nothing (0.01, $p = 0.08$). The primary school wedge also predicts the published NSC results of the school a student enters ($p < 0.001$), far more strongly than the raw average, whose 95 per cent interval includes zero. Below the leaving examination, results are not published for schools, so enrolment occurs on a directly unobserved quantity, unlike much of the school-choice literature, which measures responses to published scores (Hastings and Weinstein 2008; Abdulkadiroğlu et al. 2020).

Descriptive value added is how far a school's students achieve above students who also started with the same learning. This paper's calibrated score makes this estimable without a common

entry assessment. On the raw grade, difficult primary schools would make the secondary schools they send their students to look better than they actually are. After calibration, two in five secondary schools change quintile of value added. The reliability standard for descriptive value-added research is met: split-half 0.86 against the 0.80 benchmark (Nunnally 1978, 245; Lance et al. 2006, 206).

To my knowledge, this is the first paper to use a TWFE 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 has seen little subsequent development without a common test.

In the education literature, the TWFE 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. 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 closest design to this work is Liu et al. (2025), who also anchor school effects on the secondary school a student enters. However, a state test has already put their students on a common scale; here, the estimator itself creates the comparability across schools.

Central examination is the usual route to comparable grades, calibration is another. In Germany, an additional SD of school-leaving grades returns 6% in income 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. 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 grades cover every student but share no scale. The calibration in this paper requires no central examination or purpose-built test: the design needs no more data than education management information systems often already collect.

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 thresholds (Diamond and Persson 2016; Dee et al. 2019). The wedge is a broader object: the difficulty of a school's assessments, present in every grade it awards, far from these thresholds, as well as at them. Students in 283 schools in India achieved 64% on average on school-graded mathematics assessments. When independently retested a month later, the average dropped to 25% (Singh and Berg 2024). This paper's wedge estimates a similar gap for each primary school using secondary-school assessments.

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 tests the measure for precision and against the NSC, shows which schools set easier assessments and what cali-

bration changes in gaps and rankings, and uses it to study sorting (Section 4.8) and value added (Section 4.9). Section 5 tests the estimates against simulated threats to identification and heaping at the promotion threshold, and Section 6 sets out the uses and limits of the measure.

2 Setting and data

2.1 Grades 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 grading. 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 score is three-quarters examination and one quarter the school's own assessment. Too much time passes between Grade 7 and the NSC for the NSC to anchor the measure, but not to test it (Section 4.3).

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. Public schools teach the same national curriculum. The mathematics curriculum is a single document describing what is to be taught in Grades 7 to 9 (Department of Basic Education 2011a, foreword and §2.7).[4] Thus schools do not differ in what they are meant to teach, but may differ in the papers they set, how they grade, and how much of that curriculum they do in fact cover. The Grade 8 Term 2 grade is used because this is the term in which students simultaneously sit a grade-wide assessment. The final recorded grade used in this paper is a composite of this simultaneous end-of-term assessment and continuous assessment tasks common to all students. The Grade 7 measure is the Term 4 grade, the last report grade of primary school and the one that decides promotion. The two grades 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 *grade* throughout this paper; *score* is reserved for the calibrated common scale of Section 3 and the NSC. 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 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

[4]The 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).

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 (grades 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 grades in predictable directions. Promotion policy sets minimum grades 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.[5] 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 grades upwards across the nearest threshold.

Threshold pushing is evident in the data, as found by van der Berg et al. (2024). The grades 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 grade, which decides promotion, heaps at 40 more strongly than the Term 1 grade does up to Grade 9, and at Grade 12, whose promotion runs through the National Senior Certificate, the Term 4 excess disappears (Appendix K.1).[6]

Figure 1 shows threshold pushing in the Grade 7 mathematics data, and it sits almost entirely on one cut-off. Across the six provinces, grades are two to seven times as frequent just above 40% as just below it, and the step appears in every province and every year. Grade 8 grades show a smaller step.

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, 92% of students in the estimation sample change school between Grade 7 and Grade 8 in 2022; the rest stay on at a combined school.

In the 2022 mathematics sample of all six provinces, the median Grade 7 feeder, the primary school that sends students on, disperses its students across 7 distinct Grade 8 destinations (mean 8.6). The median secondary school receives students from 11 distinct feeder schools

[5] Department 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.

[6] These comparisons across grades and terms use a separate national extract of school-based-assessment grades from the Department of Basic Education, Terms 1 and 4 of 2021 to 2024 in all nine provinces, not the six provinces' linked records (Appendix E.2).

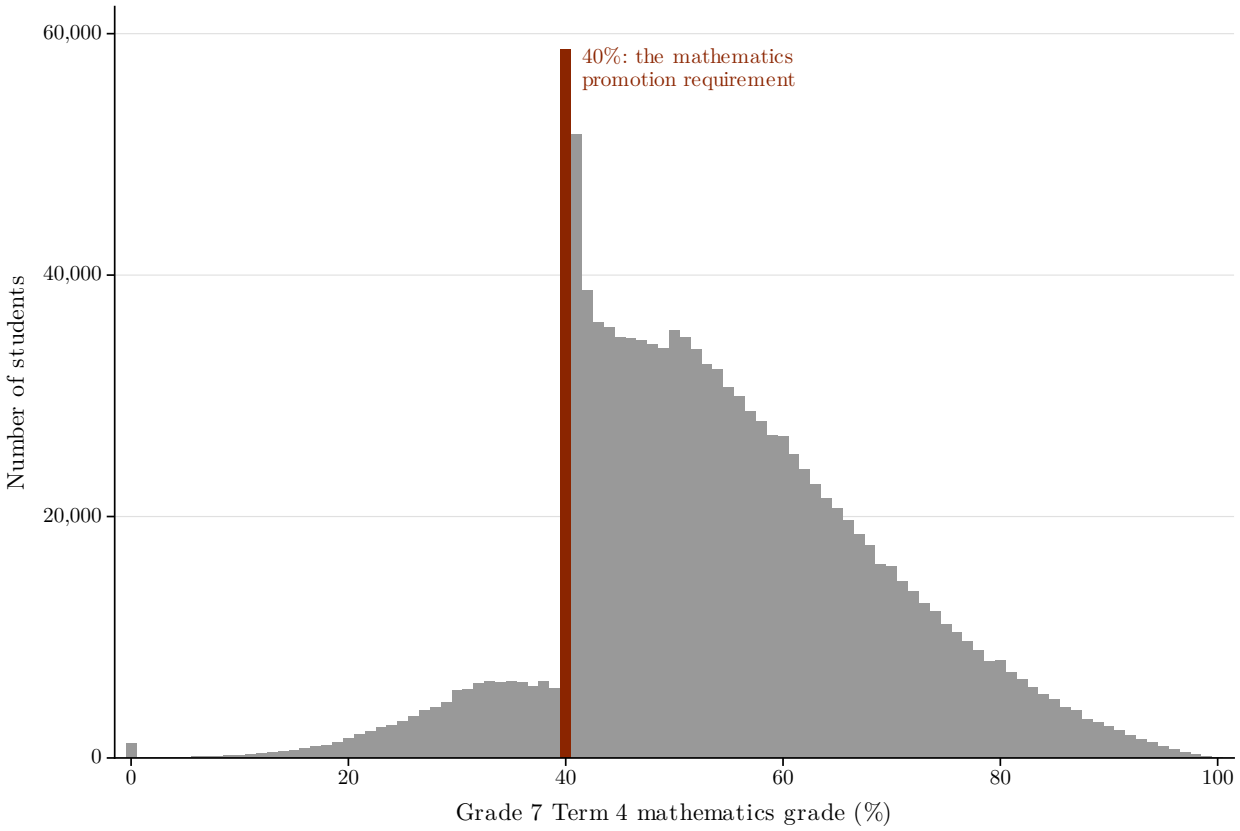

Figure 1: **Grades pile up at 40%, the mathematics promotion requirement.**

Note. Integer-bin histogram of Grade 7 Term 4 mathematics grades, 2018 and 2022 cohorts pooled. The highlighted bar is 40%. The small residual at a grade 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).

(mean 15.5). 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 hold no class identifier. The more destinations a primary school's leavers reach, the more schools that school is compared with. Gauteng's leavers are the most evenly spread of the six provinces, with a median Herfindahl index of destination shares 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. The flows in the illustrative cluster of students fan out and cross repeatedly. Feeders that never share a Grade 7 assessment are joined by many links, and destinations that never share a Grade 8 assessment are likewise joined.

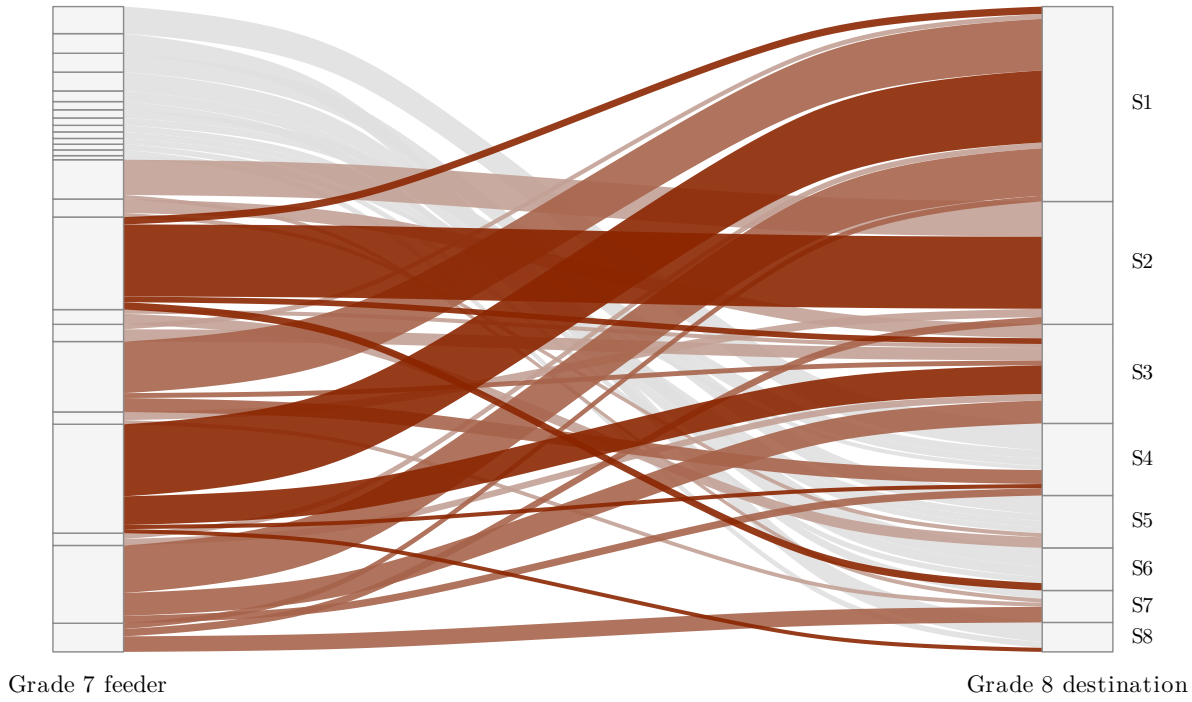

Figure 2: **Many feeders split their students across destinations, and each destination draws from several feeders.**

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.

The many-to-many mixing creates the identifying variation this paper exploits. Primaries are compared only through the assessments their leavers sit in common at a given secondary school. Because each primary feeds to many destinations, and each secondary draws from

many primaries, there is a chain of connection that joins almost all schools into a connected set. If there were no connections, however distant, between some schools and others, incomparable islands would form. Few schools fall outside: in every province and cohort the largest connected set contains at least 98.8% of primary schools (Appendix I.1), and the 103 primary schools outside hold 0.17 per cent of students.

2.4 Data and samples

Student-level records come from SA-SAMS, the administration system in which South African schools record their students' grades, 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).[7] 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 D). 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 grade, against 40% in 2018 (Department of Basic Education 2011b; Eastern Cape Department of Education 2022), so its Term 4 grade is in effect a year grade: within a school, among students with all four Grade 7 term grades, a student's Term 1 to Term 3 grades explain 94% of its variance, against 71% in 2018. The student-level test against the school-leaving examination in Section 4.3 and the value added of Section 4.9 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.

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,

[7] The sample includes independent (private) schools. Leaving out every student whose primary or secondary school was independent in the year the student attended it (3.7% of linked students 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 grade on the Grade 7 grade moves by at most 0.012. The spread of the wedge narrows most in Gauteng in 2018, to 0.92 of its full-sample SD, because the independent primaries there set harder assessments than public primaries and differ more among themselves.

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% have a valid Term 4 mathematics grade, of which 76.8% and 83.8% link to a Grade 8 record with a valid Term 2 grade. 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 I.3).

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

	2018 cohort		2022 cohort	
Step	N	%	N	%
Total Grade 7 enrolment in SA-SAMS	796,928		866,679	
With valid Grade 7 Term 4 maths grade	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.

There are two main samples used throughout the paper, clearly declared for each exhibit. The main sample is the **six provinces**. The measure is estimated separately in each province and cohort, each on its own connected set, so no two provinces' schools are ever placed on one scale. The numbers for six provinces combine 12 calibrations, computed over all schools and measured in within province-cohort standard deviations. A regression adds a fixed effect for each province and cohort. Appendix E.3 provides every number for each province. Gauteng's 106,514 students in 2018 and 135,852 in 2022 are the sample for the analyses run in one province only, such as several checks in Section 3.3, because Gauteng has the densest network (Section 2.3) and therefore the most identifying variation; each is named as Gauteng's where it appears. In the text I report the 2022 cohort: more of it reaches the sample (77% of enrolment against 67%; Table 1) and its Grade 7 grades are more reliable (0.77 against 0.70; Section 3.1). The tables report both cohorts, and the text gives both wherever a result compares them.

Table 2 describes the sample: students' grades and characteristics in Panel A, and in Panels B and C how widely primary and secondary schools mix. The original grades run from 0 to 100%; in the SD units I use, one SD is roughly 14 points in Grade 7 and 18 in Grade 8 (the

SDs of the two results across the six provinces, in both cohorts).

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

	2018 Cohort			2022 Cohort		
Variable	Mean	SD	Mdn	Mean	SD	Mdn
Panel A: Student characteristics						
Grade 7 Term 4 maths grade	50.9	13.9	49	56.8	14	56
Grade 8 Term 2 maths grade	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
Panel B: Grade 7 feeder schools						
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
Panel C: Grade 8 schools						
Grade 7 feeder schools per secondary school	12	12.4	8	15.5	14.4	11

Note. Panel A reports student-level grades 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) and 665,514 (2022).

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–grade mechanism, the IPW specification, connected-set coverage by province, and the students who do not progress to Grade 8, are reported in Appendix I.

3 Model, estimator, and identification

This section presents a model of latent learning (Section 3.1) and its estimation (Section 3.2), and explains why the comparison isolates difficulty for primary schools and not secondary schools. Section 3.3 presents what must hold for the estimates to measure difficulty alone, and the checks on each assumption. Finally, each student's calibrated score is derived from the wedge, and the characteristics of schools with easier assessments are presented (Section 3.4).

Figure 3 illustrates how the model measures the difficulty of each primary school's assessments. Two (representative) students enter Secondary 1, one from Primary A and one from Primary B, both with a Grade 7 mathematics grade of 65%. They write Secondary 1's assessments in June of Grade 8. A's student achieves 72% and B's 58% (in expectation). Had the two primary schools set equally hard assessments, the two would have scored the same at Secondary 1 (an assumption I test). Instead, their different performance at Secondary 1, despite identical primary school grades, reveals the 14-point gap in difficulty between the two primary schools' assessments, measured by each school's difficulty wedge (α_j). A set harder questions, graded them more harshly, or covered more of the curriculum. I normalise Secondary 1 as the reference secondary school ($\gamma_1 = 0$) and measure wedges against the average primary school. Equating each student's primary and secondary grades requires an additive wedge of $\alpha_A = +7$ for Primary A and $\alpha_B = -7$ for Primary B. A positive wedge denotes a primary school with harder assessments; a negative wedge, one with easier assessments. Every secondary school that receives students from both A and B repeats the comparison.

While this calibrates primary schools feeding into a single secondary school, linking across secondary schools requires students from the same primary school branching into different secondary schools. Consider a second student from Primary B (also leaving with 65%) who attends Secondary 2 and achieves 53%. Because the primary wedge is a school-level parameter fixed at $\alpha_B = -7$, equating 65% and 53% requires subtracting a further 5 percentage points. This defines Secondary 2's destination effect ($\gamma_2 = -5$), where a negative value means students score lower there than at the reference school.

In reality, each secondary school receives students from 15 primary schools on average, and each primary school sends students to 9 secondary schools; each student in Panel A represents the many who make the same move. This bipartite network density allows the model to precisely estimate all parameters simultaneously, placing every primary school within its province's connected set (at least 98.8% of primary schools) onto a common scale (Panel B). A grade on the common scale is the grade a school with assessments of average difficulty would have given the same students, so it can be compared across schools.

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 grade 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. In contrast, when two students leave Primary B with 65% and diverge to different secondary schools, they spend half of Grade 8 in different classrooms before sitting the June assessment. The 5-point gap between them

cannot separate what they learnt at each secondary school before the assessment and the difficulty of the assessments. The secondary fixed effect parameter thus combines these two components. A primary school's difficulty separates exactly because all learning from that school is reflected in the Grade 7 result they carry to their secondary school. 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 Grade 8 at secondary school k, where the Grade 8 cohort rather than the school is the unit because every student in it sits the same assessments (Section 2). a^7_{ij} is the latent learning of student i at the end of Grade 7 and a^8_{ijk} is her learning by the middle of Grade 8, both on the scale of the shared Grade 8 assessment (Section 2). Neither is observed. What is observed is the grade from each school, g^7_{ij} and g^8_{ijk}. A school's grade differs from learning in two ways: by the shift implied by the grades a school awards compared to other schools and by noise particular to that student.

$$a^7_{ij} = g^7_{ij} + s_j + \varepsilon^7_{ij}, \qquad a^8_{ijk} = g^8_{ijk} + d_k + \varepsilon^8_{ijk}. \tag{1}$$

The shift parameter s_j shows how difficult primary school j's assessments are, positive for grades reflecting more difficult assessment. Both the assessments set and how these assessments are graded compose this 'difficulty'. The secondary shift, d_k, is the same sort of object as s_j, but now for the cohort once in Grade 8 in secondary school k. The two noise terms are zero in expectation and assumed to not relate to learning or to each other. Between primary and secondary, learning is assumed to increase or decay at the same expected rate for all students in the same secondary school,

$$a^8_{ijk} = c_k + \lambda\, a^7_{ij} + \nu_{ijk},$$

where $\lambda \in (0, 1]$ is the persistence parameter, c_k is the average gain over the transition at secondary school k, 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 grades remain:

$$g^8_{ijk} = \beta_0 + \beta_1\, g^7_{ij} + \alpha_j + \gamma_k + u_{ijk}, \tag{2}$$

with $\beta_1 = \lambda$, $\alpha_j = \lambda s_j$ and $\gamma_k = c_k - d_k$ up to a constant, and an error u_{ijk} that collects the two noise terms and also how much more or less this student learned than others in her secondary school who left Grade 7 with the same learning (not grades) as she did (Appendix B gives the steps).

Equation 2 is the equation I estimate: a student's Grade 8 grade is regressed on her Grade 7 grade, with fixed effects for the primary school she left and the Grade 8 of the secondary school she entered. The primary effect, $\hat{\alpha}_j$, is the school's wedge: how hard its assessments are, scaled by persistence. It 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 graded, and anything else common to students in that grade.

If there were no noise in the grades or decay in learning between primary and secondary school, the coefficient on the Grade 7 grade would be one. Instead, it is 0.57 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 grade's reliability, ρ, which is the share of its variance that is signal after both schools' fixed effects are removed:

$$\hat{\beta}_1 \xrightarrow{p} \lambda\rho. \tag{3}$$

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 0.57 persistent and the Grade 7 grade is at least 57% signal.

A second grade for the same student separates persistence from reliability. Students who take English First Additional Language, the course for students whose home language is not English, also have a Grade 7 English grade. It comes from different assessments, so its noise is largely unrelated to the noise in their mathematics grade, while it tracks learning at the same school. The slope on the part of the mathematics grade that the English grade predicts is free of the mathematics grade's noise, so it measures persistence alone: about 0.74 in 2022 (Appendix B.12). 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 grade's reliability: about 0.77. Noise and fade share the shortfall roughly equally, with fade a little the larger. Both grades 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 grade sits on the right-hand side of the regression, its noise attenuates the slope, and through the slope it reaches the wedge:

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

up to a constant, where $\bar{a}^7_j$ 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 grade, 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 grade, removes both distortions. The spread of the wedges widens from 0.453 to 0.527 within-province SDs of students' Grade 8 grades in 2022. The gap between no-fee and fee-charging primary schools widens from 0.41 to 0.47 SD; Section 4.5 takes it up. However, the two sets of wedges are correlated at 0.992, which means that the ranking of primary schools is almost unchanged. Differences in student intake alone account for only 10–12% of the variance in the wedge and 13–14% 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 G.5).

3.2 Estimation

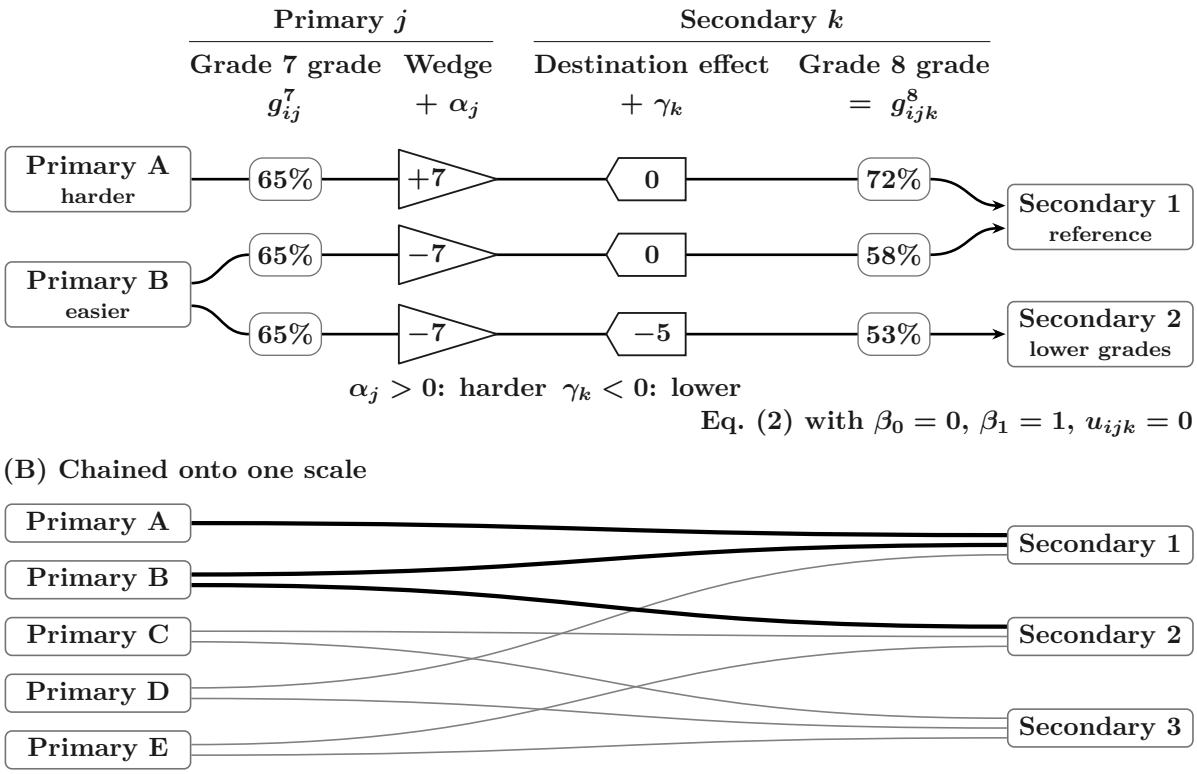

Figure 3: **Same grade in, different grades out: the gap between primary schools measures difficulty.**

Notes: The grades are illustrative. Panel A: each row is one student. The sum of her Grade 7 grade, her primary's wedge, and her Grade 8 cohort's destination effect gives her Grade 8 grade, 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 grade 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. In the data, $\hat{\alpha}_j$ captures the average extent to which primary school j's leavers score higher or lower in Grade 8 than their Grade 7 grades 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 grades 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.[8] 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.

I estimate Equation 2 by ordinary least squares with high-dimensional fixed effects (Bergé et al. 2026) on grades standardised within province and cohort. Section 4.8 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 0.96 or more in both cohorts (Appendix G.1). Table 3 reports the estimates. The wedges spread widely. Their SD across schools is 0.453 SD in both cohorts, 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 grade, can land more than a full SD of Grade 8 grades apart. Each province shows much the same spread (Section 4.1, Appendix E). The destination effects spread more: their SD across Grade 8 cohorts is 0.598 SD in 2022.

This paper reports results in two units, Grade 7 standard deviations and Grade 8 standard deviations. All calibrated results, the wedges and the destination effects are in Grade 8 SDs. The raw primary grades and a student's own grade are in SDs of Grade 7 grades. Because learning attenuates between the two grades, a gap needs to be rescaled by that attenuation in order to equate Grade 7 SDs to Grade 8 SDs. 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 grades at the same secondary school. The convention in this paper is that whenever a raw grade is compared to a calibrated one, the raw grade is rescaled by the persistence parameter, about 0.75.

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

	2018	2022
Panel A. Specification		
Students	532,295	665,019

[8]The destination effects are identified only relative to one another, as the wedges are, and the fit is unchanged if a constant moves from every wedge to every destination effect. A wedge is never read against a destination effect, which is early teaching net of difficulty, not difficulty.

(continued)

	2018	2022
Origin (Grade 7) schools	10,978	11,890
Destination (Grade 8) cohorts	7,002	6,637
Slope on the Grade 7 grade	0.541	0.573
Standard error (two-way clustered)	(0.004)	(0.003)
Adjusted R-squared	0.635	0.600
Within R-squared	0.334	0.354
Panel B. The wedge		
Standard deviation	0.453	0.453
Median standard error	0.100	0.098
Different from the average school at 5% (%)	62.3	63.8

Notes: Ordinary least squares estimates of Equation 2 in each cohort, with fixed effects for the primary school and the Grade 8 cohort. Grades are standardised within province and cohort, so the slope is in within-province SD of the Grade 8 grade per SD of the Grade 7 grade, and Panel B's first two rows are in SD. The slope's standard error is two-way clustered by primary 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:*** *the six provinces, mathematics, 2018 and 2022 cohorts, each province fitted on its own; the slope is the mean of their slopes.*

The wedge does not rank students, it is not a causal effect of the school, nor does it reward a school based on the students it admits. A's wedge of +7 is a school-level effect, and grading noise averages out as the number of students at the school grows. The comparison is between students with the same Grade 7 grade, so a stronger intake increases a school's grades, not its wedge, apart from the small intake part that noise adds (Section 3.1) and any learning not measured by the Grade 7 grade (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 grade at a given secondary school, whichever primary they left, and that their Grade 7 grades 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 grade a student as the two firms: the feeder in Grade 7 and the destination cohort 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 effect is not a fixed effect but rather her Grade 7 result, measured with noise, which is why $\hat{\beta}_1$ estimates $\lambda\rho$ and why noise reaches

the wedge (Section 3.1). The setting has two advantages over labour markets: 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 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), Appendix A gives their detail, and Appendix B.9 lists every condition formally.

The first assumption is that, for a given level of learning (not grades) when a student exits primary school, which school she left makes no further difference to her grades. 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:

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

The data cannot test this directly. Going back to the first comparison of Figure 3: from the same Grade 7 grade, 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 grade shows the same gap under either story, so I maintain the assumption.

Whether A's students learn faster after leaving can be checked in part. Because Grade 8 has four terms of grades, 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 grades of students from primary schools with harder assessments increase by 0.026 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. In Gauteng, Term 4 grades follow the Grade 7 grade about a tenth more closely than Term 2 grades do, so a primary school's wedge shows up about a tenth larger by Term 4: in the example, A's 14 points would become about 15 with no extra learning at all. Across the six provinces, difficulty alone predicts a gain of 0.037 SD, more than the 0.026 observed, so students from primaries with harder assessments learn no faster during Grade 8 than their peers (Appendix G.2). 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 grades, and differ between primary schools, to bias the results.

The same limit applies in the other direction: the estimator cannot separate incomplete curriculum coverage from lenient grading. Unlike test equating, statistical moderation uses an anchor assessment to rescale each school's grades. 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 assessments and the anchor measure the same thing (Masters 1988) (Appendix C). Here the shared Grade 8 assessments are the anchor. The regression compares only averages and interprets the gap between a school's grades and its students' results 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 graded 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. Under either interpretation, a grade 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 grade more strictly.

The second assumption, AKM's exogenous mobility, is that among students from the same primary school with the same Grade 7 grade, which secondary a student enters is unrelated to how much she learns over the transition and to any learning her result missed. Formally, given her Grade 7 grade and her primary school, the secondary she enters is independent of her error in Equation 2 (Appendix B.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 grade 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 grade, whether based on their rate of improvement or on dimensions of learning that the Grade 7 grade 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 H.7 allows that sorting but imposes persistence of one.

Students do sort into destinations, but mainly on the Grade 7 grade the regression conditions on; sorting on gender and over-age is weaker, though beyond chance (Appendix H.2). Sorting on how fast students improve would show before the move, because Grade 7 has four terms of grades: had B's leaver bound for Secondary 2 been improving more slowly, her grades would already have risen less over Grade 7 than those of B's leaver bound for Secondary 1. In Grade 7 grades, 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: within 0.012 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 0.453 SD of Grade 8 grades (Table 4 gives Gauteng's, with standard errors). This is the pre-trend test of exogenous mobility (Card et al. 2013, 983–86). In Gauteng in 2018 students sort a little on transitory grade shocks, which leaves the order of the wedges almost unchanged (Appendix H.1).

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

Assumption: the threat to it	Check (where reported)	Result	Reading
Panel A. Standard two-way fixed-effects checks			
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 I.1)	At least 98.8% of primary schools in every province-cohort sample; 99.4% [99.4%] and 99.7% [99.7%] overall; the leave-one-out set keeps 10,864 [1,421] of the 10,978 [1,424] and 11,830 [1,472] of the 11,890 [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 I.3)	Noise is 16.1% [5.4%] and 13.4% [5.5%] of the wedge's variance; SD 0.415 [0.458] and 0.421 [0.455] corrected, against 0.453 [0.471] and 0.453 [0.468]	Small, not nil
Exogenous mobility: students bound for different destinations were already on different paths before the move	Grades of students who change school over the four Grade 7 terms, by destination, within the same feeder (Appendix H.1)	Cohorts pooled: split halves [0.001 (0.005)]; other cohort [0.012 (0.006)]; mean grade [0.027 (0.006)]	Near flat by destination effect; a few hundredths of an SD by destination mean grade

(continued)

Assumption: the threat to it	Check (where reported)	Result	Reading
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 × destination decile; a slope per destination (Appendix J.1)	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 mean residual [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 H.2)	Cohorts pooled: excess over chance, Grade 7 grade 16.9, gender 4.1, over-age 5.8 percentage points	Mainly on the conditioned grade
Additivity: a feeder does not grade all its students by one shift	Mean residuals by the student's Grade 7 decile × wedge decile; a slope and shift per Grade 7 quintile (Appendix J.1)	Largest absolute mean residual, net of the Grade 7 margin shared by all feeders, [0.369] and [0.374] (low-grade 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 grade	Size varies with the grade; order barely moves under a slope per feeder

(continued)

Assumption: the threat to it	Check (where reported)	Result	Reading
Additivity: a school grades some groups of its own students differently	Separate wedges by student group within the same schools (Appendix J.2)	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 grades
The student's own effect: conditioning on the Grade 7 grade stands in poorly for each student's own effect	A student fixed effect from the eight term grades (Appendix H.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 F.2)	The wedge correlates 0.477 [0.624] between the 2018 and 2022 cohorts, precision-weighted (9,918 [1,280] schools)	Bears on use, not identification
Panel B. Specific to school-set grades			
Exclusion restriction: a feeder's effect on the Grade 8 grade is not all difficulty: the feeder taught skills its grades missed, or its students grow differently after the move	The exclusion restriction (Appendices G.1 and G.2)	Cohorts 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. Magnitudes are in within-province SD unless a row names another unit, and the exogenous-mobility row's are in SD of Grade 7 term grades; rank correlations are Spearman's, other correlations Pearson's. Cohorts are estimated separately, 2018 first, unless a row says cohorts pooled. Numbers in square brackets are Gauteng's; a row that prints only a bracketed number reports that check for Gauteng alone. Standard errors and the joint test are clustered by primary school. Each check uses its own sample, defined in Appendix A.1. ***Sample:*** *the six provinces, mathematics, 2018 and 2022 cohorts (532,295 and 665,019 students; 10,978 and 11,890 primary schools); in square brackets, Gauteng*

(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 2.3). Sampling error inflates the estimated variance of school effects under sparse linkage between schools, the limited-mobility bias documented in 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 13.4% of the wedge's variance in 2022 (5.5% in Gauteng, where the network is densest; 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' grades by one amount. Net of the Grade 7 margin shared by every feeder, low-grade students at the feeders with the easiest assessments land 0.324 SD above the model's prediction in 2018, against a wedge SD of 0.453 (Card et al. 2013, 995–96; Bonhomme et al. 2026, 19) (Appendix J.1; Table 4 gives Gauteng's). The size of B's wedge, −7 in the example, therefore depends on which of its students one looks at. Its rank does not. In Gauteng, 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 A). Across destinations the effects do add: in Gauteng, a separate effect for every pair of feeder and destination adds little to the adjusted R^2, though the pairs are jointly significant in 2018 (Table 4).

The cardinal size of wedges is more exposed than the rank of wedges. In Gauteng, the placebo of Appendix G.4 shows how socioeconomic status inflates $\hat{\alpha}_j$, so the wedge still partly reflects selection. Part of that is the intake component added by noise (Section 3.1). In the Monte Carlo simulations, both sorting and endogenous mobility distort size but not rank (Section 5.1).

Two threats are particular to this setting. First, grades are pushed over the 40% promotion threshold (Section 2.2) in both primary and secondary. Section 5 tests what this does to the estimates. Second, within a school, different classes may be graded differently. Basic fairness norms mean that schools set the same assessments for all the students in a grade, but this does not rule out teacher-specific grading. The wedge remains unbiased so long as this teacher effect does not correlate with the schools students arrive from. There are no class identifiers, so this condition is assumed.

3.4 The calibrated score, and the shape of the wedge

To get a student's calibrated score, their place on the common scale, their Grade 7 grade is multiplied by the slope, plus the school's wedge. The opening's two students both have a Grade 7 grade 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 grade below 30% land on average 0.17 SD higher in Grade 8 than their calibrated scores say, and students with 80% or more land 0.18 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 grades 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 grade net of destination effects: where its leavers land.

The wedge varies more across schools than the level does: an SD of 0.453 against 0.368 SD in 2022. 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

$$\hat{\alpha}_j \approx \overline{g^8}_j - \hat{\beta}_1 \, \overline{g^7}_j \tag{4}$$

which equates the wedge to a school's mean Grade 8 grade, net of the destination effects, less a $\hat{\beta}_1$ fraction of its mean Grade 7 grade. Where two schools' leavers do equally well in Grade 8, the school with the lower Grade 7 grade 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.

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-grade quintile, so easier assessments are not equivalent to weakness. The two cuts run in opposite directions. Difficulty falls with the raw grade, 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 in Gauteng the gradient also survives a Grade 7 slope fitted for each poverty quintile and a later Grade 8 term (Appendix G.2); Section 4.5 takes it up.

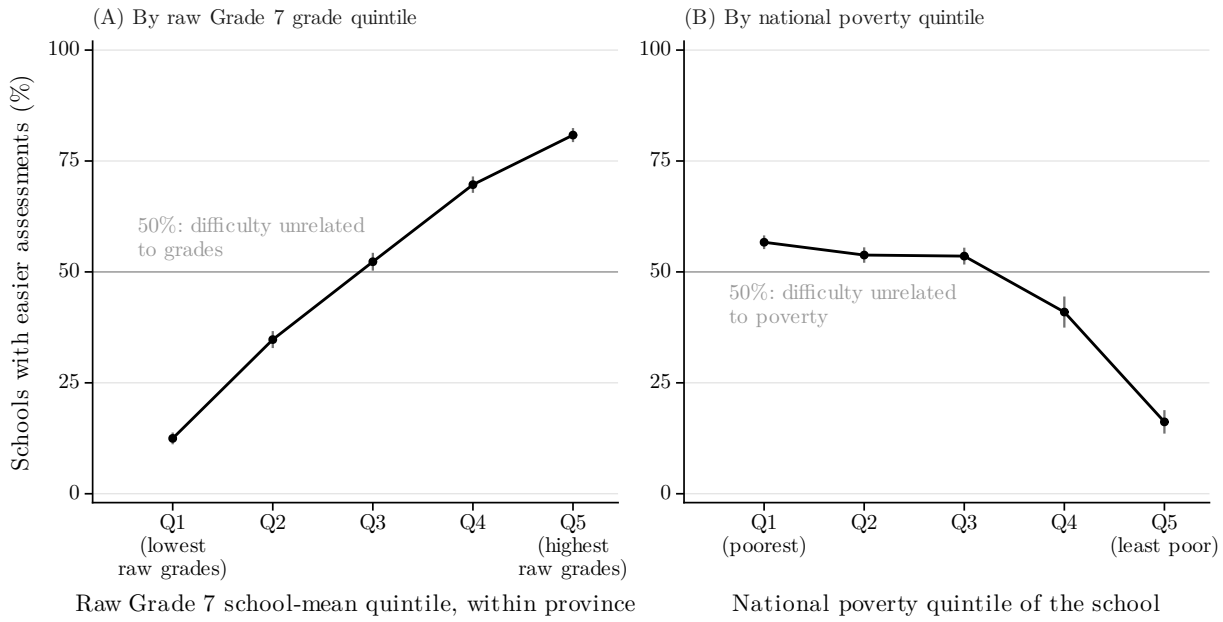

Figure 4: **Schools with easier assessments appear at every raw-grade quintile, and the richest schools have the hardest.**

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.*

4 Results

This section shows how much the wedge varies (Section 4.1), how precisely it is estimated (Section 4.2) and whether it predicts the NSC (Section 4.3). It then shows how calibration changes results estimated using raw grades: the poverty gradient (Section 4.4), correlates of easier assessments (Section 4.5), gaps across cohorts (Section 4.6), and the ranking of schools by performance (Section 4.7). Finally, two applications show what the calibrated score tells us about sorting into secondary schools (Section 4.8), and how it might be used for value-added estimation (Section 4.9).

4.1 How much schools differ

Primary schools differ widely in how hard their assessments are. The SD of the wedge across primary schools is 0.45 in both 2018 and 2022. This is nearly half the within-province SD of students' Grade 8 grades: about 8 points out of 100, compared with about 18. 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 grade, one from each of

these schools, would therefore score about 20 points apart in the same secondary school. The dispersion is similar in every province. Estimated separately for each of the six provinces in each cohort, the SD of the wedge ranges from 0.43 to 0.49.

The variation within a district is larger than the variation between districts. Of the wedge's variance, 94.4% is within districts (precision-weighted). Two schools differ almost as much whether randomly drawn from within a district as from within a province. If there were district-level moderation or rubrics, this would likely decrease observed within-district variation in wedges.

4.2 Precision and persistence

The wedge is estimated precisely. In estimating reliability, I use the Kane–Staiger signal share (Kane and Staiger 2002). Each school's wedge is the average, over its students, of how far a student's Grade 8 grade lies from what her Grade 7 grade and her secondary school predict. The students' estimands are estimated with noise. The larger the school, the more the noise cancels. The variance in the wedge thus has two parts, the real difference between schools and the school-level estimation noise that remains post aggregation. Reliability is simply the first part as a share of the total. This share is 0.87 in 2022 (0.95 in Gauteng, where the network is densest), above the 0.80 the literature recommends for research (Nunnally 1978, 245–46; Lance et al. 2006, 206). Estimation noise accounts for about 13 per cent of the differences in the wedge between schools (full decomposition in Appendix F.1).

A more demanding check asks, if a school's students are randomly split in half, with the estimation run independently for each half, whether those two halves correlate highly. If the wedge were a function of particular students rather than the school, these halves would correlate poorly. Instead, even with the intrinsic increase in noise, the two halves agree closely: averaged over the six provinces, the split-half correlation is 0.743 in 2022 (0.879 in Gauteng).

The wedge is also not sensitive to the choice of estimator or to student demographics. Other estimators (quadratic, piecewise-linear, empirical-Bayes, and cross-classified) order schools alike, each correlating with the baseline at 0.964 or above (Gauteng's in Online Appendix Table 29). Re-estimating the calibration with over-age, gender and population-group controls leaves the ranking essentially intact: the re-estimated wedges correlate with the baseline at 0.998 or above in both cohorts. In Gauteng, wedges estimated separately for over-age and on-age students, or for girls and boys, also order schools alike (Appendix J.2).

The wedge persists between the 2018 and 2022 cohorts, with a correlation of 0.477, and the calibrated learning level at 0.712 (0.624 and 0.889 in Gauteng); the raw Grade 7 mean persists less, at 0.386 (Table 5). Net of estimation noise, about 52% of the wedge's variance recurs in the other cohort, about 45% is specific to the cohort though shared by all four Grade 8 terms, and about 3% is specific to the Grade 8 term (Appendix F.2). Even though the COVID-19 pandemic fell within those four years and nothing links one year's scale to the next, this level of stability suggests that the estimate reveals something about the school itself.

Schools still change rank substantially between cohorts: 65.9 per cent land in a different quintile of calibrated learning. 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 correlations and unity mixes cohort composition, drift in assessment difficulty, estimation noise, and change in the feeder-to-destination network (Appendix F.2; Appendix F.5).

Table 5: **A school's 2018 calibrated level predicts its 2022 level far better than its raw mean.** *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 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 0.477 across the same schools (Appendix F.2); the calibrated learning level's Spearman correlation is 0.465.* ***Sample:*** *the six provinces, 9,918 schools present in both cohorts, each province fitted on its own; every measure is centred within its province and quintiles are formed within province.*

2018 measure	2022 raw Grade 7 mean	2022 calibrated learning level
Raw Grade 7 mean	0.386 (71%)	0.241 (77%)
Calibrated learning level	0.163 (78%)	0.712 (66%)

The wedge also reproduces across subjects. A school's English and mathematics wedges move together: disattenuated for estimation error (8,732 schools), they correlate at 0.601.

4.3 Validation against the National Senior Certificate

In the absence of a contemporaneous common assessment in Grade 7 or Grade 8, the school-leaving examination, the NSC, validates the calibration. I can link the NSC for one cohort: students in Grade 7 in 2020, to their results in 2025. In each of the six provinces and across them together, the calibrated score predicts students' NSC results better than the raw grade does. The correlation rises by about a quarter in NSC Mathematics, from 0.313 to 0.389, and by about three-fifths in Mathematical Literacy,[9] from 0.248 to 0.398. Additionally, at the same Grade 7 grade, students from primary schools with harder assessments do better in the NSC. Students with a raw Grade 7 grade in the 60s score 5.7 points more in NSC Mathematics, on average, if their primary school is in the third of its province's schools with the hardest assessments rather than the third with the easiest (Figure 5).

The calibrated score's advantage is not an artefact of the secondary schools students enter. At the same Grade 7 grade, students from primary schools with harder assessments enter secondary schools with higher-achieving peers (Section 4.8), so the advantage could come from those schools. Within the same secondary school, the calibrated score still predicts

[9]Students 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).

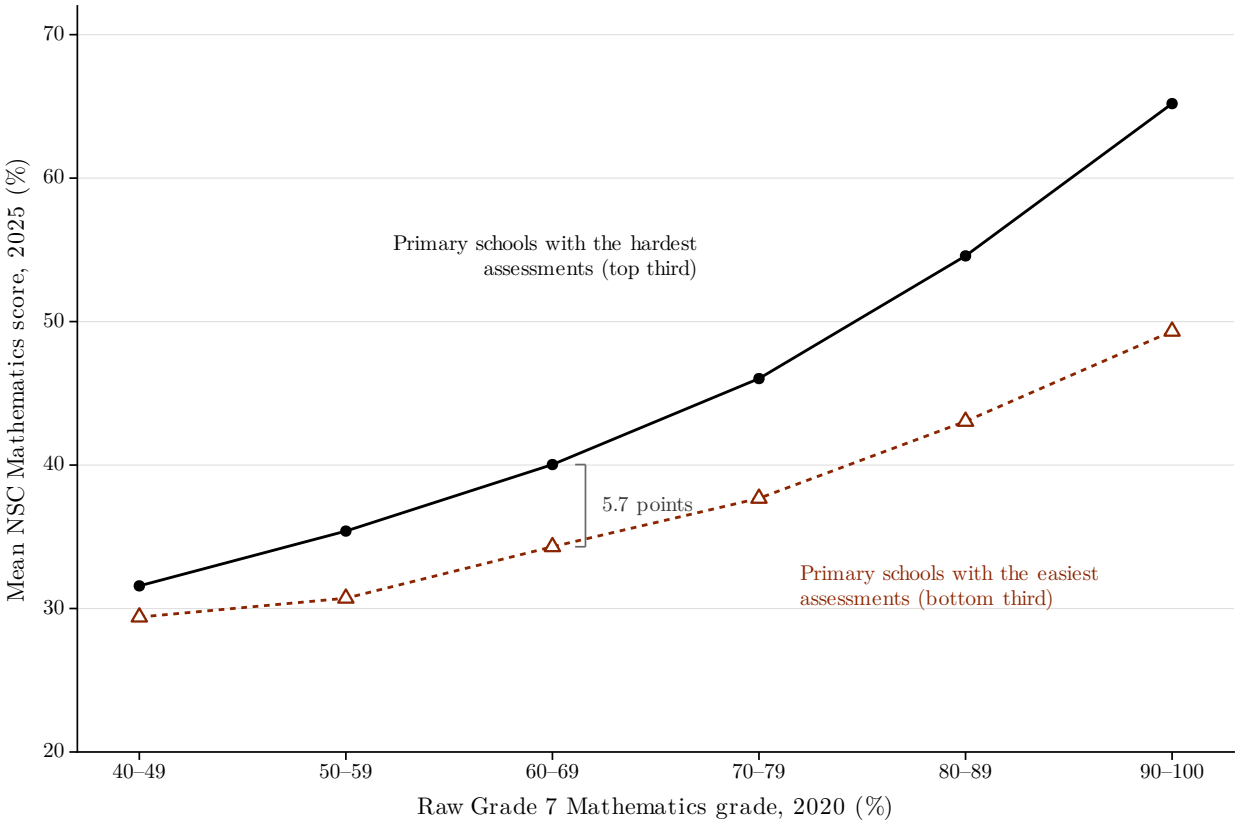

Figure 5: **At the same Grade 7 grade, harder primary assessments predict higher NSC Mathematics scores.**

Notes: The comparison does not hold the secondary school fixed; Figure 7 does. Each point is the mean NSC Mathematics score in 2025 of the students whose raw Grade 7 Term 4 Mathematics grade in 2020 falls in a ten-point band, for students from the third of their province's primary schools with the hardest assessments by the 2020 wedge and from the third with the easiest. The band below 40% is not drawn: for the easiest third it holds 16 students who sat NSC Mathematics. The NSC score is out of 100, a quarter school-based assessment and three quarters external examination; Figure 7 measures it in national SDs. The share of each province's 2020 panel linked to the NSC is lowest in Gauteng, 38%, against 64–78% in the other five provinces. Sample: the six provinces, the Grade 7 2020 cohort who sat NSC Mathematics in 2025, 26,007 students from the third with the hardest assessments and 19,278 from the third with the easiest.

the NSC score better than the raw grade, in each subject, in every province and across the six provinces (Figure 6). The within-school R-squared rises from 0.152 on the raw grade to 0.191 on the calibrated score in Mathematics, and from 0.160 to 0.204 in Mathematical Literacy.

Among students who entered the same secondary school with the same raw Grade 7 grade, 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 predicts a probability of sitting the NSC higher by 0.340, from a base of 0.396, against 0.196 for a one-unit higher own Grade 7 grade (Figure 7, Panel B). Among those who sat, the wedge predicts the student's own NSC score at 0.697 national NSC SDs 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 grade (Panel A; Online Appendix Table 37). All of these estimates are net of the origin school's quintile and no-fee status, and of its rural status in the four provinces whose school register records it.

These estimates have three limitations; Appendix M.1 discusses the first. First, only a fraction 0.396 of students sat both the Grade 7 assessments and the NSC five years later. As the wedge predicts whether a student sits the NSC at all, the sample is selected on the regressor, an effect I do not bound. Second, students who sat both assessments are not balanced on observables: the NSC records link to 75.6 per cent of students from quintile-1 primary schools but 23.9 per cent of those from quintile 5. The estimates therefore lean towards poorer primary schools where assessments are easier (Section 4.5). Third, the wedge is estimated with noise, and the standard errors here are estimated as if it were known, understating uncertainty (Appendix L.1).

4.4 What a production function recovers

Replacing the raw mean with the calibrated level as a dependent variable in a static learning production function changes estimated gradients. I regress the raw mean Grade 7 grade and the calibrated learning level separately on poverty quintile (from Quintile 1, the poorest, to Quintile 5), log enrolment, log class size and the student–teacher ratio, with province-by-cohort effects. The sample is the 11,977 of the 22,868 school-cohorts for which every characteristic is recorded. Online Appendix Figure 13 reports, for Gauteng, 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 students are older than the age expected for the grade. Over-age is excluded from the first regression because assessment difficulty in part determines whether a student is held back and therefore over-age. Across the six provinces, in the second regression the coefficients on the three richest quintiles, log enrolment, log class size, the student–teacher ratio and over-age differ significantly between raw grades and the calibrated level: the 95 per cent bootstrap interval of each difference excludes zero. That on the second quintile does not.

The shifts are also large, and the largest is in the socioeconomic gradient. The Quintile 5 advantage rises from 0.24 SD under the raw mean to 0.68 under the calibrated level, so raw grades show only 36 per cent of it, about a third. Raw grades understate rich schools'

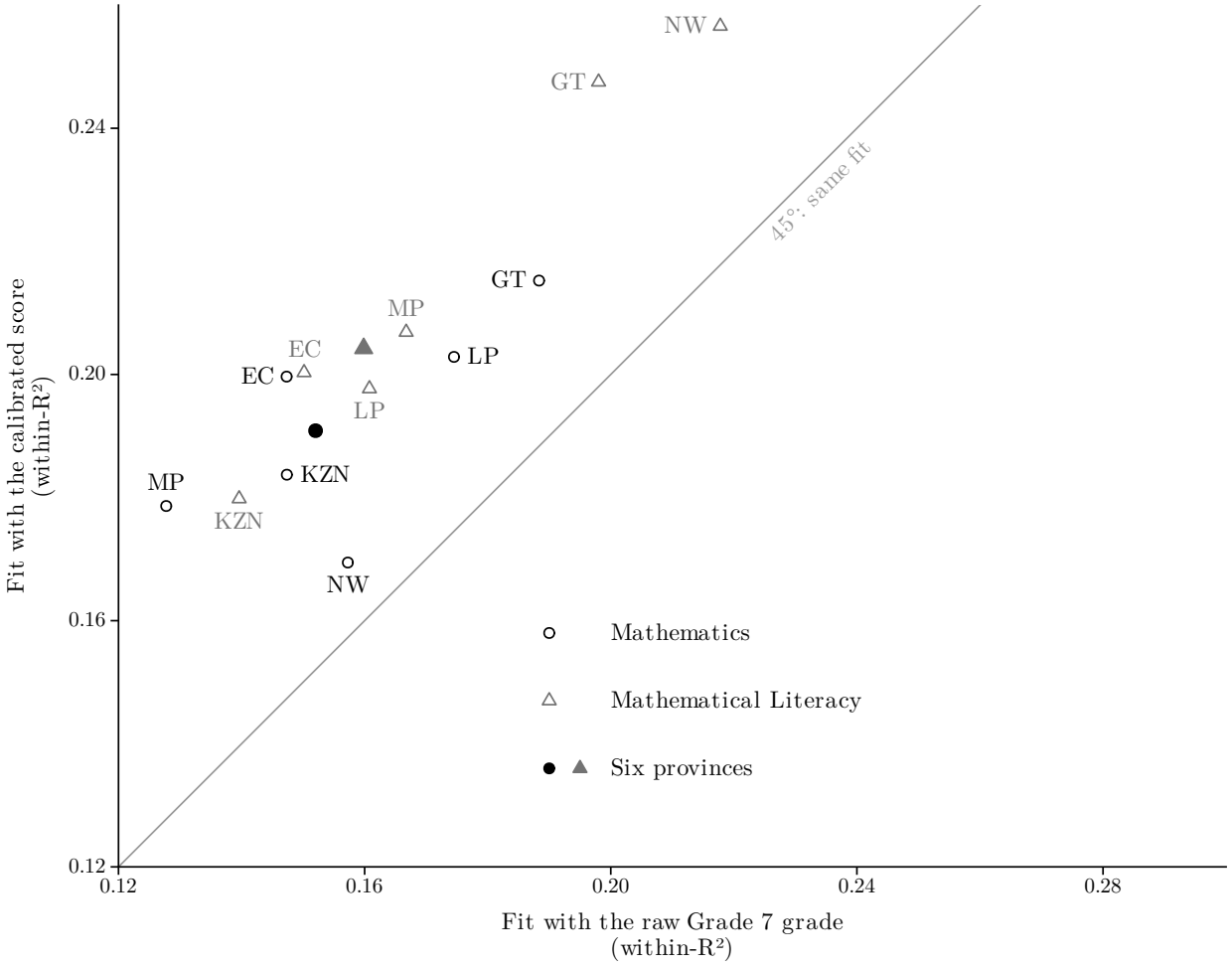

Figure 6: **Within secondary schools, the calibrated score predicts the NSC score better than the raw grade.**

Notes: Each point is a regression of the student's NSC 2025 score on her Grade 7 grade, with a fixed effect for the secondary school she entered and the same controls throughout, for one subject in one province or across the six provinces. Open marks are single provinces, labelled EC (Eastern Cape), GT (Gauteng), KZN (KwaZulu-Natal), LP (Limpopo), MP (Mpumalanga) and NW (North West); filled marks are the six provinces together. The horizontal axis gives its within-R^2, the share of the variation in NSC scores within secondary schools that the regression explains, with the raw Grade 7 grade; the vertical axis gives it with the calibrated score. A point above the 45-degree line is one where the calibrated score explains more. No weight in the calibrated score was fitted to the NSC, and the two regressions are not nested, so the comparison is of fit, not a formal test. Sample: the Grade 7 2020 cohort in six provinces, 70,790 students who sat NSC Mathematics in 2025 and 85,454 who sat Mathematical Literacy.

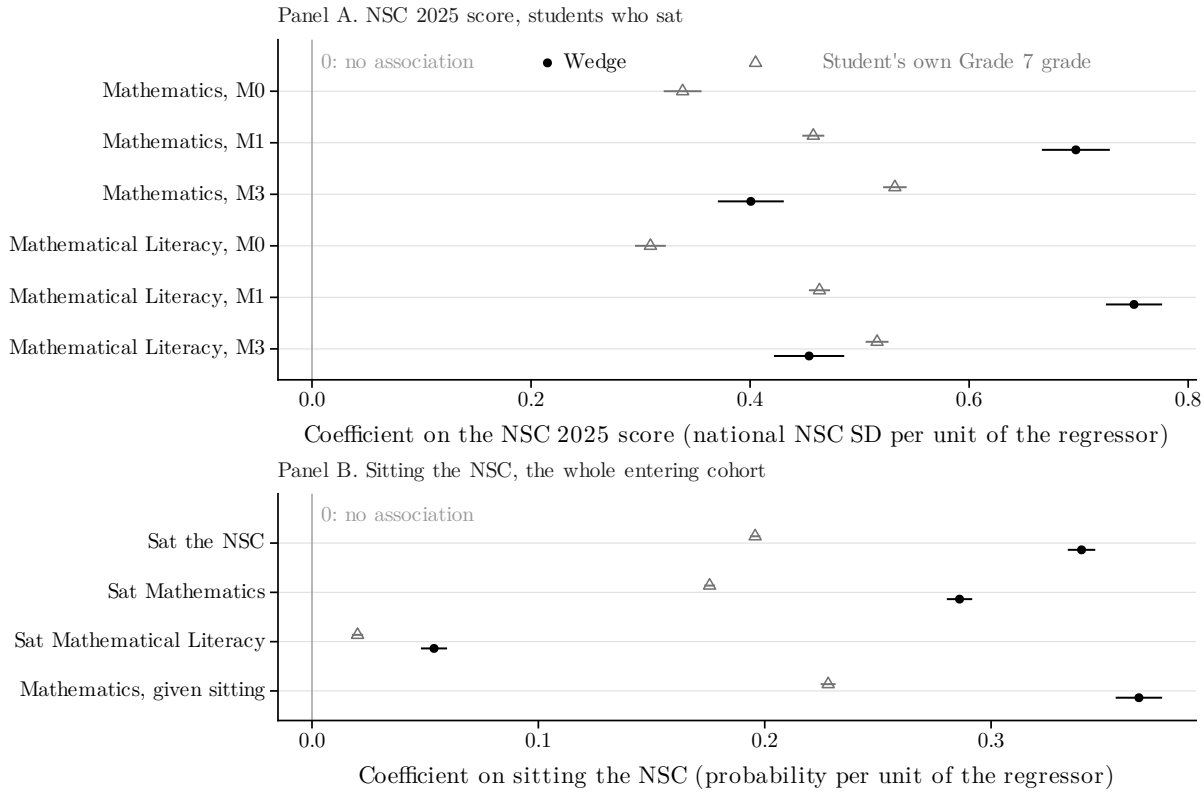

Figure 7: **Same secondary school, same Grade 7 grade: harder primary assessments predict better NSC results.**

Note. Each row is a regression with a fixed effect for the secondary school of entry and controls for sex, home language, population group, over-age, and the origin school's quintile, no-fee status and rural status (not recorded in Mpumalanga or North West). M0 has the own Grade 7 grade alone, M1 adds the wedge, and M3 adds the origin school's mean Grade 7 grade, a component of the wedge, so the wedge's fall under M3 is built in. M2, which adds the origin school's mean grade to M0 without the wedge, is not shown. In Panel B, the last row is taking Mathematics among the students who sat the NSC. Coefficients are per unit, not per SD of this sample, and the two regressors' units differ, so their order does not rank them. Whiskers are 95 per cent intervals, clustered by origin school. **Sample:** the 394,353 students of the Grade 7 2020 cohort who entered a secondary school in one of six provinces with a calibrated intake grade (Panel B), of whom 70,790 sat NSC Mathematics and 85,454 Mathematical Literacy (Panel A).

students' learning most (Figure 8). The coefficient for class size also moves: it grows from −0.07 to −0.13 SD per log point, with the confidence interval of the change outside of zero.

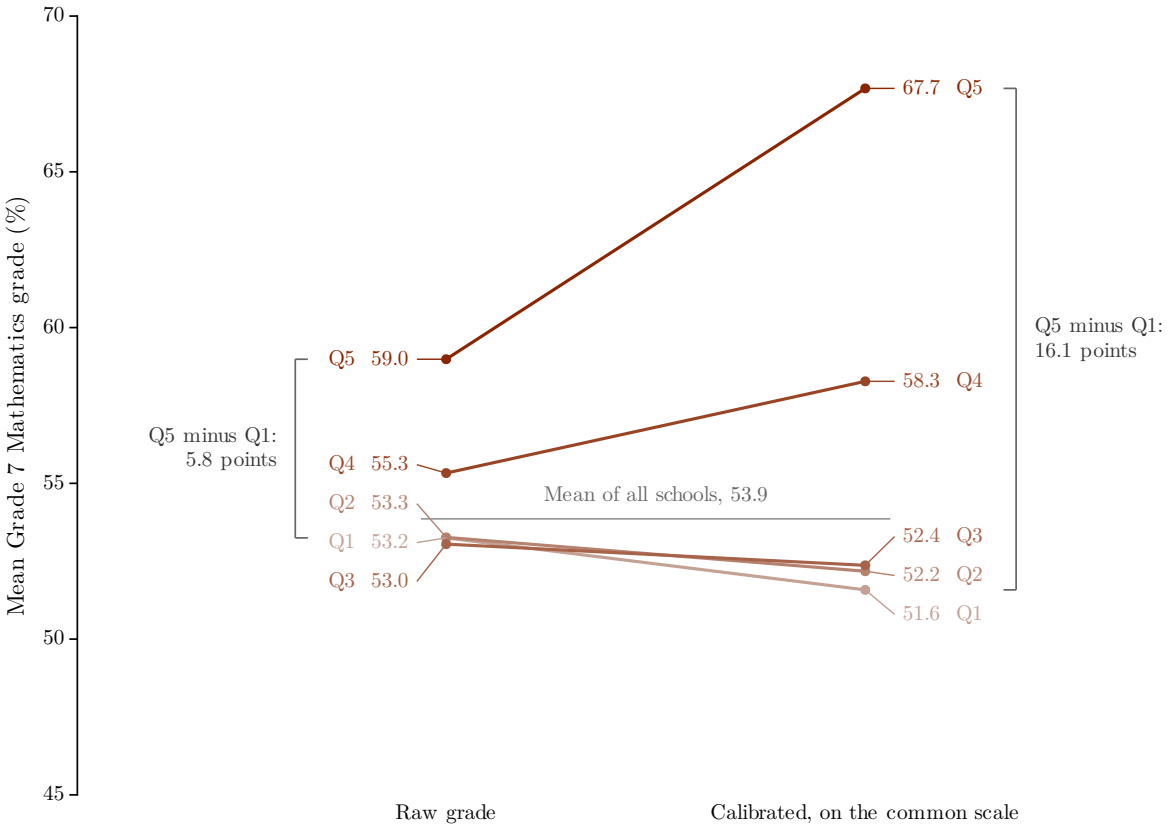

Figure 8: **Raw grades hide most of the gap between the richest and the poorest primary schools.**

Notes: Each line joins the mean Grade 7 Mathematics grade of the primary schools in one poverty quintile on raw grades (left) to their mean on the common scale (right), with no controls. On the common scale a school's grade is its raw mean plus its wedge expressed in Grade 7 grades: the grade a school in its own province with assessments of average difficulty would have given the same students. Each school's level is converted within its own province and cohort, so no school is put on another province's scale. The wedges are centred within each province and cohort, so the mean of all schools is the same on both sides. One SD of Grade 7 grades is roughly 14 points. Sample: the six provinces, 2018 and 2022 cohorts, 22,845 school-cohorts; schools whose Grade 7 grades were not captured are excluded.

Allowing for estimation error in the calibrated level does not change any coefficient's significance. As the calibrated level is estimated, standard errors clustered only at the primary school level would understate uncertainty. Instead, I use a paired cluster bootstrap with $B = 1{,}000$ replicates. Schools are resampled, so a school's 2018 and 2022 observations move together, and the wedge is redrawn within every replicate before re-estimation. Gauteng's bootstrapped intervals in Online Appendix Figure 13 are not substantially larger than the clustered ones, because the wedge is estimated precisely (Section 4.2).

4.5 Which schools set easier assessments

Poverty reliably separates schools with easier assessments from schools with harder ones. The estimates in this section are Gauteng's: the regressions control for whether a school is rural, which the school register does not record in Mpumalanga or North West. Among the school's resources and structure, two covariates stay significant even after a multiple hypothesis testing correction (a Romano–Wolf stepdown). They are the no-fee indicator, at an adjusted p below 0.001, and the school's own poverty quintile, at 0.008. Both show that poorer schools have easier assessments.

On its own, the no-fee indicator's standardised coefficient is 0.490; with the school's own poverty quintile, its other resources and the mechanical controls held fixed, the coefficient is 0.426 (Figure 9; Appendix L.2). 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. Across the six provinces, where rural status cannot be held fixed, the no-fee indicator's coefficient on its own is 0.320.

If the wedge measured differential learning decay between primary and secondary rather than assessment difficulty, the covariates' coefficients would change depending on the term used for estimation. Instead, the coefficient on no-fee status, re-estimated on Term 2, Term 3 and Term 4 grades, is essentially flat (Appendix G.2). Likewise, if the wedge measured strategic inflation at passing grade thresholds, the schools that heap at these thresholds would show as having easier assessments, the opposite of what is in fact found (Section 5.2).

37.1% of the variance in the persistent wedge is explained jointly by the covariates. What I label the "downstream market structure", the secondary schools a primary school's leavers enter, explains the most, 38.3% of the explained variance, but no single family of covariates explains the pattern in isolation (Online Appendix Table 36).

4.6 The rich–poor gap across cohorts

The change in the gap between rich and poor schools over time is also distorted by raw grades. Figure 10 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. On raw Grade 7 grades, 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.[10] 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. 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).

[10] A cohort's results are standardised within its province and year. The estimator does not link cohorts, so these gaps are relative to each cohort's spread. This allows comparability between cohorts if the students' learning spreads about as widely in each cohort. A change in the variance would move the raw and calibrated gaps in the same direction, but it could not make the raw gap halve while the calibrated gap remains the same.

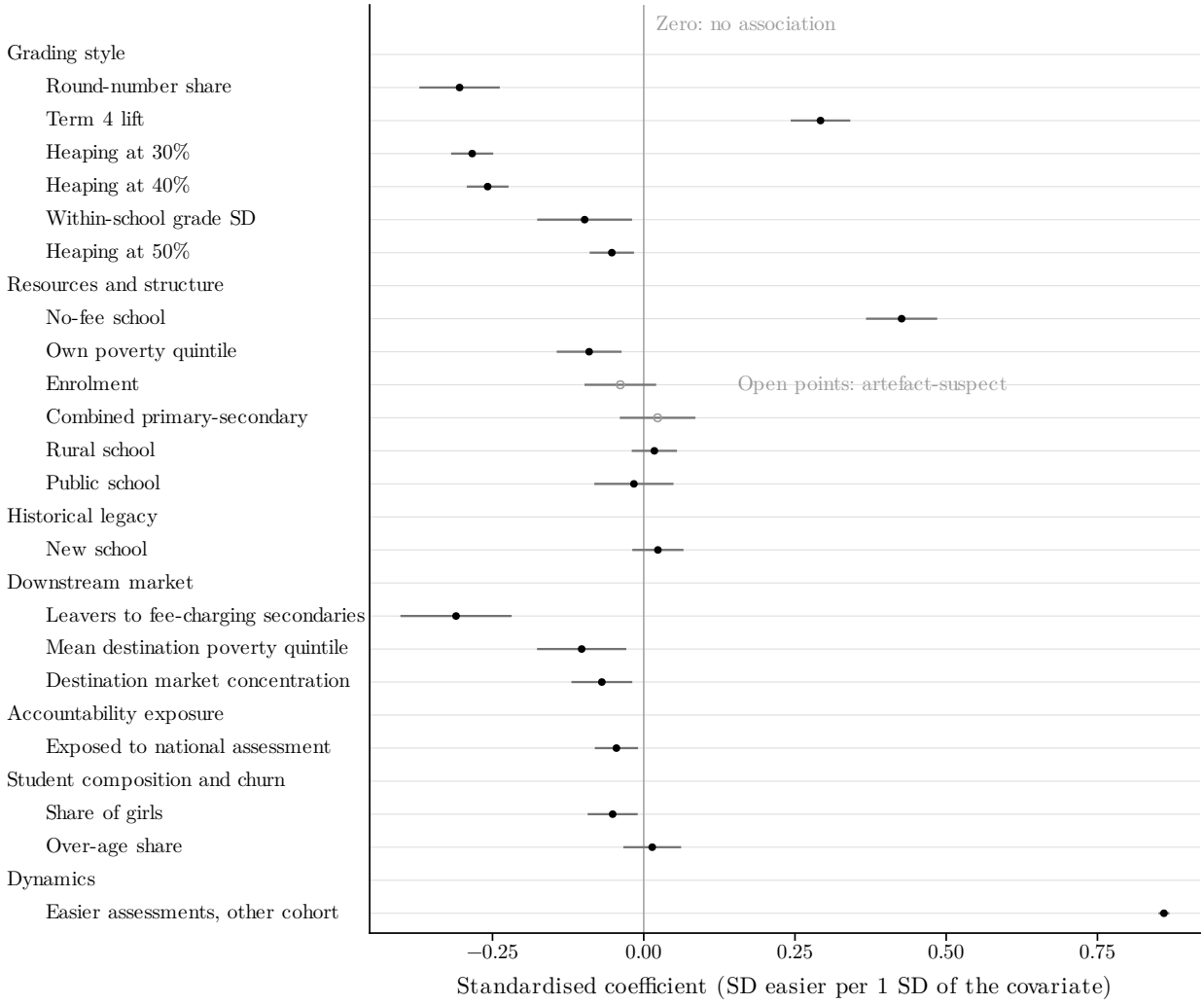

Figure 9: **In Gauteng, no-fee status has the largest association among school characteristics with easier assessments.**

Notes: 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 SDs of the outcome per SD of the covariate. The outcome is the sign-reverse of the wedge, so a positive coefficient is easier assessments. Horizontal bars: 95% confidence intervals. Term 4 lift is the school's mean Term 4 grade less its mean over Terms 1 to 3; round-number share is the share of Term 4 grades divisible by 5; exposed to national assessment marks a school that took part in the Annual National Assessments; and a new school is one founded after 1994. The open points, enrolment and combined primary–secondary schools, are the two covariates whose association with the wedge largely disappears once the mechanical family is held fixed, so they may track how precisely the wedge is estimated rather than the school. 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 (Appendix L.5). Associations only; no causal interpretation. Appendix L.2 gives the full covariate atlas and the variance decomposition. Sample: Gauteng Mathematics, up to 2,753 pooled school-cohorts with an own poverty quintile (1,280 schools seen in both cohorts; precision-weighted).

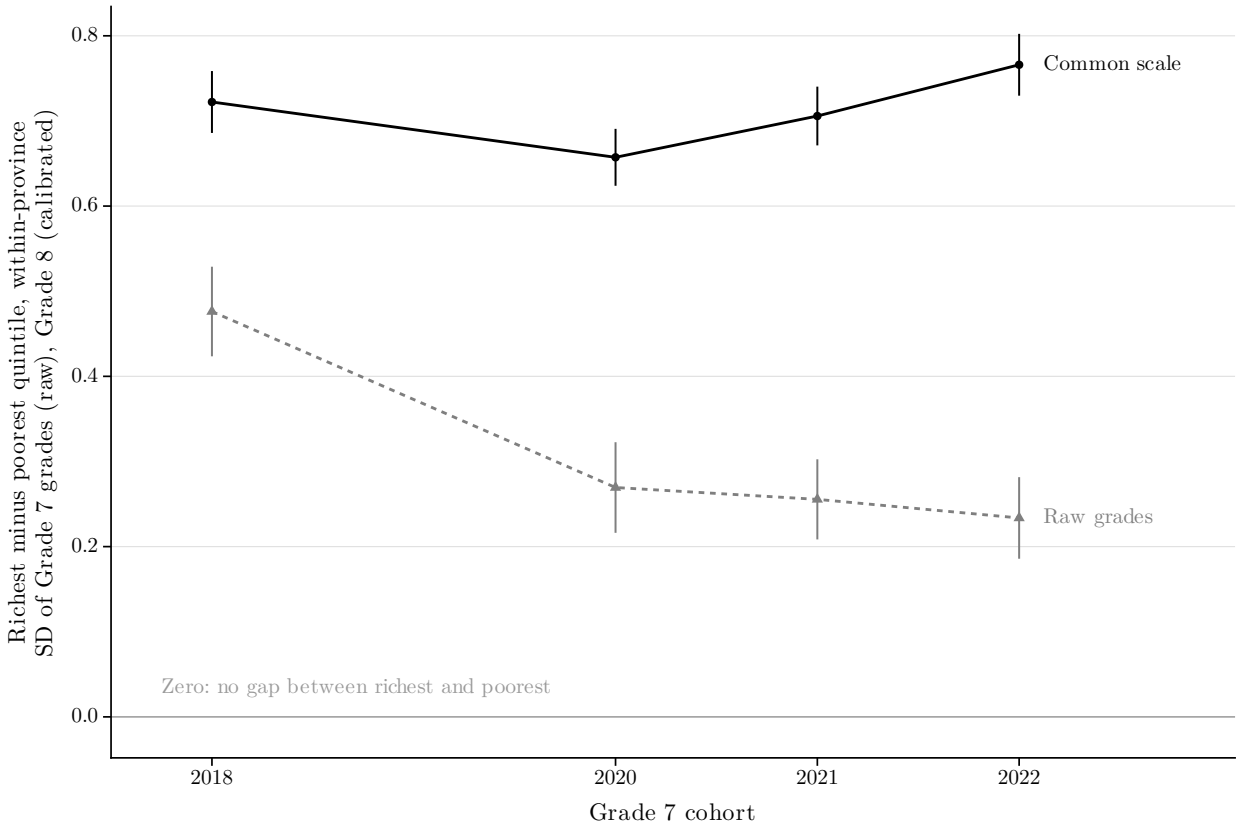

Figure 10: **The rich–poor gap roughly halved on raw grades and held on the common scale.**

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 grades to second-term Grade 8 grades. Raw gaps are in within-province SD of Grade 7 grades, calibrated ones of Grade 8 grades. Whiskers are 95 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 grades. Sample: the six provinces; 10,396 to 11,299 primary schools a cohort.

4.7 Which schools change rank

Put on the common scale, most schools change quintile: 77.1% of the 22,868 school-cohorts are in a different quintile of calibrated learning from the one their mean Grade 7 grade puts them in (Figure 11). No quintile on raw grades keeps even half its schools. Only 22.4% of the schools in the bottom quintile on raw grades are in the bottom quintile of learning, so support aimed at the bottom quintile on raw grades would miss more than three in four of the schools whose students have learned least. At the top, 29.4% stay. In the middle, where schools are packed most closely and a small change in level is a large change in rank, only 20.4% stay, and the rest move up and down in about equal numbers. In all, 44.7% of school-cohorts move two quintiles or more.

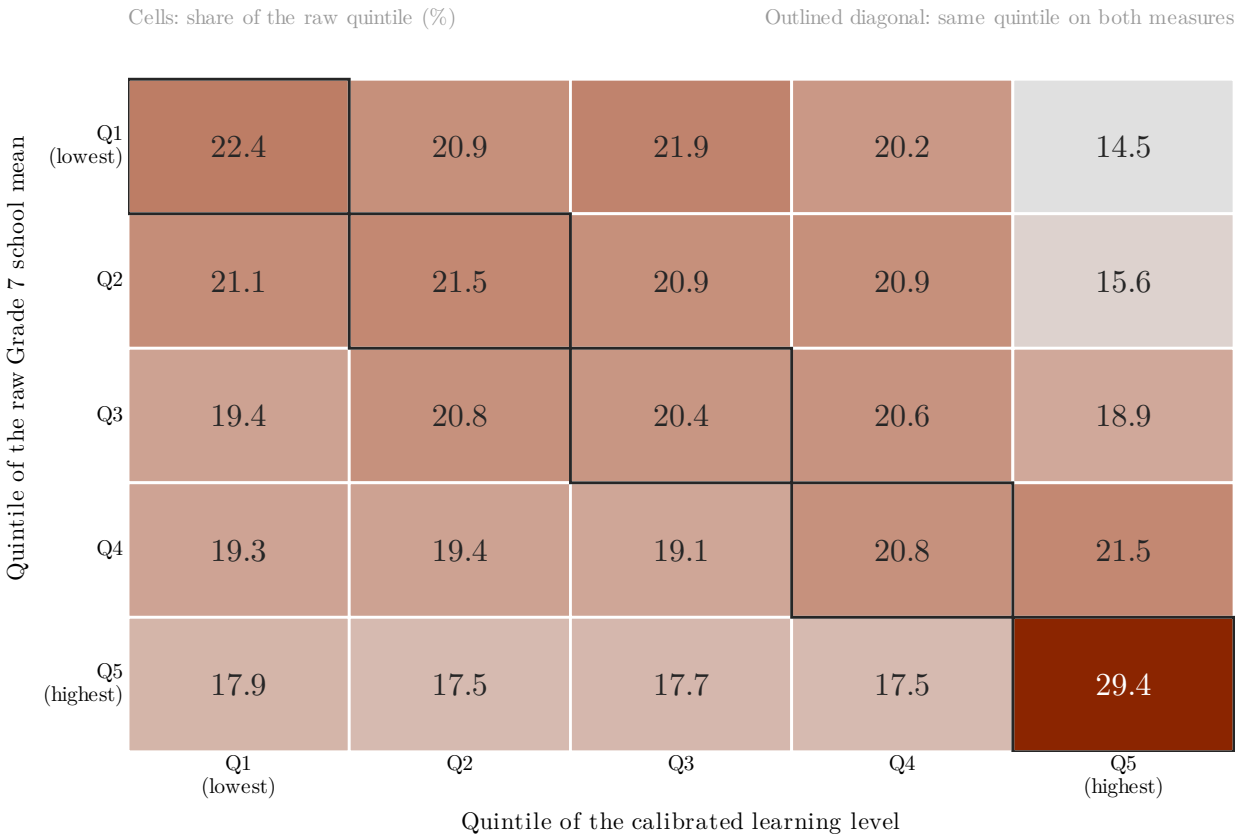

Figure 11: **Calibration moves schools out of their raw-grade quintile throughout the distribution.**

Notes: Each cell is the share (%) of the school-cohorts in a quintile of the raw mean Grade 7 grade (rows) that fall in each quintile of the calibrated learning level (columns), with quintiles formed within each province and cohort and all school-cohorts counted together. The outlined diagonal holds the school-cohorts in the same quintile on both measures; 77.1 per cent lie off it. Rows may not sum to 100 because of rounding. Sample: the six provinces, 2018 and 2022 cohorts, 22,868 school-cohorts.

The re-ranking is not driven by small schools, whose wedges are the noisiest: within each province, the number of a school's leavers (on a log scale) explains at most 0.9% of the

variation in how far it moves.

4.8 Sorting into secondary schools

Students who leave primaries with harder assessments join higher-achieving Grade 8 cohorts, holding fixed their own grades. 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. No result is published for primary schools, so the signal that households are acting on must arrive indirectly, such as through siblings, reputation, local networks, etc.

I test for sorting by regressing the mean grade of the Grade 8 cohort a student joins against variables of her primary school, one at a time: the Grade 7 mean or wedge, holding her own Grade 7 result and demographics constant. A student's Grade 8 peers are composed of those who arrived from her primary school and from other primary schools. The students from her school may carry primary-school signals, so I also run the sorting regressions leaving out her primary school feeder peers. This leave-own-feeder-out (LOFO) specification does not change the pattern observed. Across the six provinces, the wedge's LOFO coefficient is about half its value in Gauteng, where it is largest (Online Appendix Table 42).

The wedge of a student's primary school, not its raw mean grade, predicts how well her new Grade 8 peers do. A student from an origin whose wedge is one SD higher reaches a Grade 8 cohort whose Term 2 grades are 0.117 SDs of the peer mean higher, and 0.064 higher excluding her primary-school peers (0.136 in Gauteng, as Figure 12 shows). The raw mean grade predicts nothing: its coefficient is +0.012 under LOFO (+0.020 in Gauteng), within sampling error of zero. This test does not show whether households are making decisions in light of how hard their school's assessments are.

I check three other explanations that might produce this result (Appendix N). First, the wedge partly contains the student's Grade 8 result, so it might pick up her unobserved ability. Replacing her cohort's wedge with the same school's wedge from the adjacent year's cohort returns a coefficient of 0.068 ($p < 0.001$). Second, if each secondary school's effect is subtracted from the outcome, the coefficient rises from 0.064 to 0.230 ($p < 0.001$). This suggests that students from harder primaries are not systematically going to secondaries with easier grading. Third, there may be geographic sorting on assessment difficulty, schools with harder assessments feeding 'stronger' secondary schools. In Gauteng, the coefficient holds among only those who cross a ward or municipal boundary. Nonetheless, the data cannot separate families who choose a school from families who move to be closer to one (Bayer et al. 2007; Rothstein 2006).

At the same own grade, students from primary schools with harder assessments enter secondary schools with stronger published NSC results. Unlike any primary-school result, a secondary school's NSC results are published, and reported in the press every January. Across the six provinces, pooling the two cohorts, a student from an origin whose wedge is one SD higher enters a secondary school whose published NSC Mathematics mean is +0.104 SDs higher ($p < 0.001$). Her primary school's raw mean grade predicts −0.009 (with 95 per

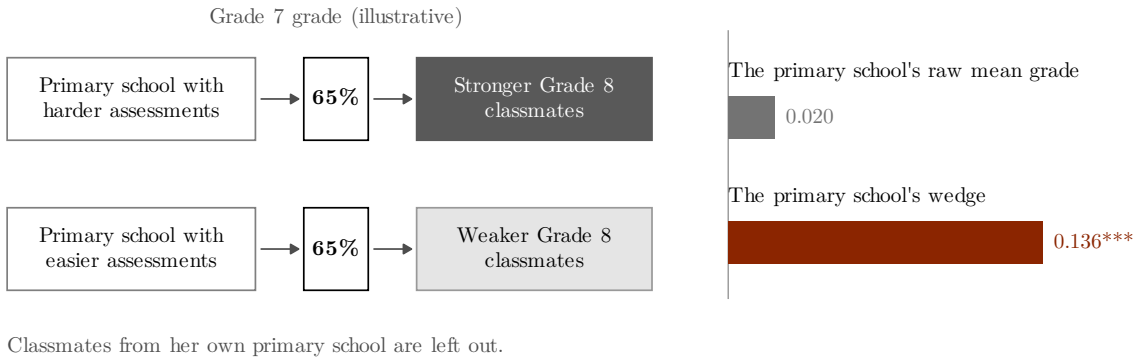

Figure 12: **In Gauteng, at the same grade, students from primaries with easier assessments join weaker peers.**

Notes: The diagram illustrates the comparison; the grade of 65% is illustrative, not data. The bars report the coefficients on two signals of the student's primary school, its raw mean Grade 7 grade and its wedge, each from a separate estimate of the regression in Appendix N.2 that holds her own Grade 7 grade and the controls fixed. The outcome is the mean Grade 8 Term 2 grade of the cohort she joins, leaving out every peer from her own primary school (column 4 of Online Appendix Table 39), in SD of that mean per SD of the signal. Standard errors are clustered by primary and by secondary school. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. *Sample: Mathematics movers, Gauteng, 2018 and 2022 cohorts, the leave-own-feeder-out sample of Online Appendix Table 39, n = 220,973.*

cent interval −0.022 to 0.005); the wedge predicts more strongly (0.088 to 0.119), and the two intervals do not overlap. As the NSC results of the years before entry are earned by cohorts at least 5 years before the sampled cohort, there is no inclusion of her primary school peers. Rather, this reflects whom a school admits alongside what it teaches. The results do not show that the chosen school teaches better.

4.9 Value added

A school's value added is the difference in results between where a school's students ended and where they started. This aims to separate what a school adds from whom it admits. Comparability is thus required for both the entrance assessment and the exit assessment. With the calibrated score as the intake, secondary schools' effects on the school-leaving result become descriptive value added.

Secondary schools differ widely in value added. For student i entering secondary school k, I estimate

$$Y_i = \theta_k + \delta A_i + X_i'\varphi + \varepsilon_i \tag{5}$$

separately in each province, on the same 394,353 students as Section 4.3. Y_i is her NSC score, standardised nationally; A_i is her calibrated score; and X_i holds the controls of Section 4.3. θ_k is the exit-school effect. It shows how far school k's entrants finish above or below what their entrance assessments and controls predict, relative to the province's average secondary school. Every statistic below covers the 1,135 (1,486) secondary schools with at least twenty

sitters of the subject, Mathematics first and Mathematical Literacy, estimated separately, in parentheses. Net of estimation noise, the effects' SD is 0.344 (0.321): students with the same intake at two secondary schools one SD apart finish about a third of a national SD apart.

The effects estimated with the raw and with the calibrated intake agree at a disattenuated rank correlation of 0.877 (0.867), yet 40.4% (44.2%) of secondary schools change quintile between them (Mathematics in Figure 13). On raw grades, a student from a harder primary school arrives in a secondary school with a grade below her learning, and would then finish above what that grade naively predicts. Thus, estimating value added on raw grades would attribute differences in primary school difficulty to secondary schools.

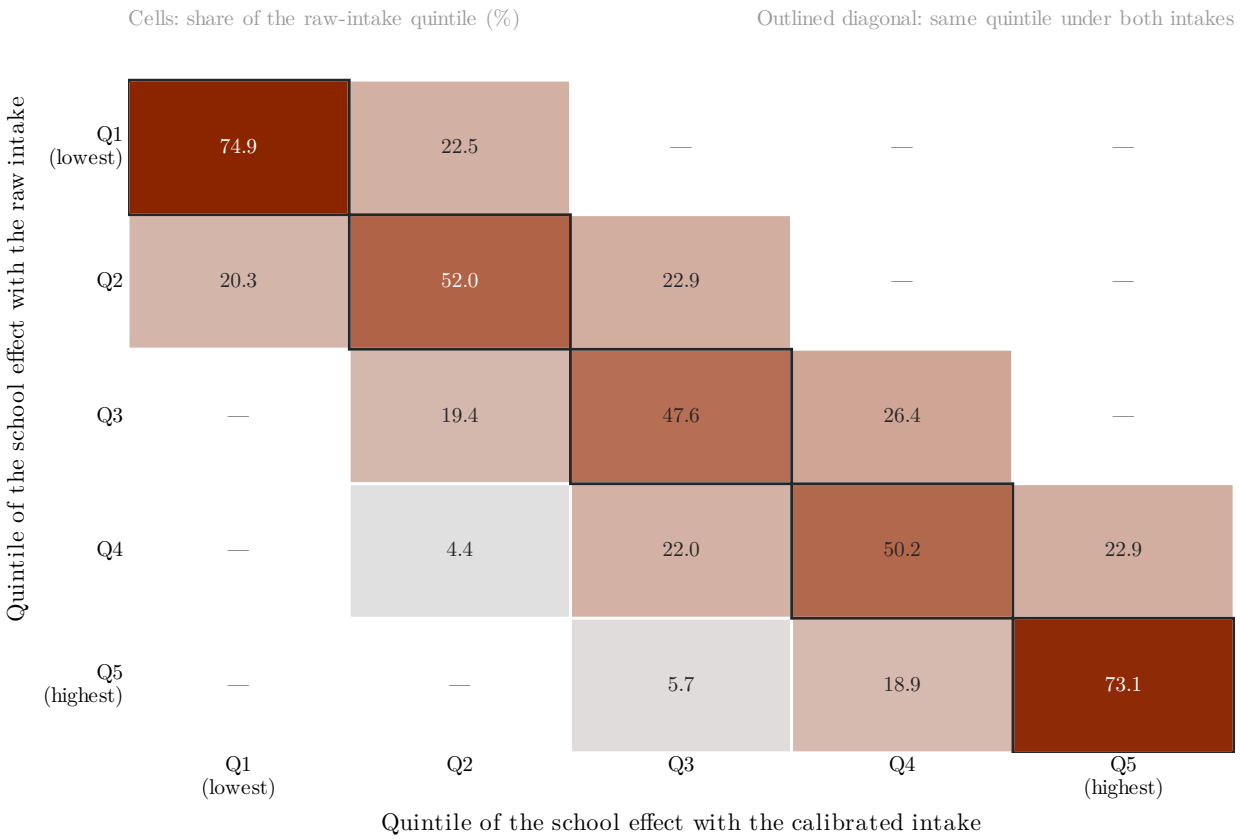

Figure 13: **Two in five secondary schools change their quintile of value added with the calibrated intake.**

Notes: Each cell is the share (%) of the secondary schools in a quintile of the Mathematics effect θ_k estimated with the raw Grade 7 grade as the intake (rows) that fall in each quintile of the effect estimated with the calibrated score as the intake (columns). Both sets of effects are centred within province. The outlined diagonal holds the schools in the same quintile under both intakes; 40.4% lie off it. A cell of fewer than ten schools is suppressed and printed as an em dash, so rows may not sum to 100. Sample: the cohort in Grade 7 in 2020 and the NSC in 2025, in six provinces, the 1,135 secondary schools with at least twenty Mathematics sitters.

The raw intake fails the standard placebo test for an intake, and the calibrated intake passes it. The test re-estimates the effects with a grade earned before entry as the outcome. No secondary school can have changed its entrants' Grade 7 grades, so school "effects" on such a grade that disperse as widely as θ_k mean that the effects record who enrols. The placebo is the Grade 7 Term 1 Mathematics grade, three terms older than the Term 4 grade the intake is built from. With the raw Grade 7 grade as the intake, the raw Term 1 grade disperses across secondary schools at 0.696 (0.704) of the effects' own dispersion, above one half. With the calibrated score as the intake, the calibrated Term 1 score disperses at 0.364 (0.434), under half.

Sorting that the calibrated intake misses is small against the effects' signal: the calibrated Term 1 Mathematics placebo accounts for 1.8% (2.8%) of the effects' signal variance. Adjusting θ_k for this sorting leaves 95.1% (95.1%) of secondary schools in the same band, above, indistinguishable from or below the province average. The proportional-selection bound of Oster (2019), which scales unmeasured sorting with measured sorting, leaves 78.0% (77.7%) there (Online Appendix Table 52).

The effects are precisely estimated: split-half reliability is 0.855 (0.841), above the 0.80 the literature recommends for research (Nunnally 1978, 245–46; Lance et al. 2006, 206). The effects are most sensitive to attrition, the students who entered a secondary school but have no examination result. Under the median of the adjustments for it, the adjusted and unadjusted effects still have a disattenuated rank correlation of 0.861 (0.801) (Online Appendix Table 52).

Reliability only shows that θ_k is a school trait rather than noise, not that it is learning. A school's grading is such a school trait. A quarter of each student's NSC score is set and graded by her school, but Umalusi, the education quality council, bounds these grades to within a narrow band of the NSC examination results (Umalusi 2016). Thus, generous grading can only raise θ_k and reliability within that band.

Causal effects are not tested in this paper. However, comparable entrance assessment is a necessary condition for causal value added, should the setting also have as good as random admission (Deming 2014; Angrist et al. 2017). A forecast test on a second cohort, per Chetty et al. (2014), will arrive when the 2021 Grade 7 cohort writes the NSC in 2026.

5 Robustness

This section shows that the main threats to the estimates, the sorting of students into schools and classes, thin or endogenous mobility and grades pushed over the promotion threshold, leave the ranking of schools largely in place. In simulated data with known school effects, the ranking of schools holds under sorting and mobility, though sorting and endogenous mobility distort the size of the estimates (Section 5.1). Heaping could be mistaken for easy assessments, but in Gauteng schools with harder assessments heap more, not less (Section 5.2). The ranking also holds when heaping is simulated at its observed strength and when the observed grades are aggressively smoothed (Section 5.3). Selective attrition is answered with the data in Section 2.4: a school's linkage rate is nearly unrelated to its wedge.

5.1 Simulated threats to identification

Robustness to a set of the most likely threats to identification is tested using Monte Carlo simulations on a synthetic school system with Gauteng's network density. Data are generated with known school effects under three threats: sorting, thin mobility between feeders and destinations, and endogenous mobility. The ranking survives every threat. At Gauteng's network density, and with endogenous mobility as strong as the data suggest, the rank correlation with the known school effects holds at or above 0.95. It falls to 0.88 only when each primary sends 15 leavers to two secondary schools, and to 0.90 under the strongest endogenous mobility simulated (Online Appendix Table 21).

Sorting acts on the spread of the estimates but leaves their order largely intact. Take the example of Section 3, in which Primary A's assessments are 14 points harder than Primary B's. If A's leavers go mostly to Secondary 1, a school whose grades come out higher, no harm is done: the destination effect for Secondary 1 takes that advantage, so A is not credited with it. But a secondary school teaches its Grade 8 cohort in several classes, and the data do not record a student's class, so the destination effect is one number for the whole cohort. If the classes students land in are unrelated to the primary school they came from, this does not bind. Only sorting into classes based on primary school origin breaks this. In the simulation, the leavers from harder primary schools are placed in higher-performing classes, so the regression credits A with part of those classes' advantage, and the gap between A and B widens beyond the underlying 14 points, "stretching the ruler". At the strongest sorting simulated, the variance of the estimates is 2.4 times the true variance, while their rank correlation with the truth rises from 0.96 to 0.98.

In the opposite direction, endogenous mobility narrows the spread. The data show that this effect is likely smaller than matters. In the simulation, the abler students within each primary, who also have the higher Grade 7 grades, reach the classes where grades come out higher. The regression reads their better Grade 8 grades as a steeper return to the Grade 7 grade, nets too much of each school's average Grade 7 grade out of its wedge, and the gap comes out narrower than 14 points ("shrinks the ruler"). The strength of this sorting is measured between secondary schools: across the six provinces, the within-feeder correlation between a student's Grade 7 grade and her destination's estimated effect is 0.021. That correlation would have to reach 0.048 in the simulation at Gauteng's network density, 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 H.8).

5.2 Heaping at the promotion threshold

The Grade 7 grades that enter the calibration are themselves manipulated: grades bunch at 40%, the promotion requirement in mathematics, and hardly at 30% or 50%, which promotion does not require of mathematics (Section 2.2; Online Appendix Table 30). A school can give a student a higher grade for the same learning in two ways. It can grade leniently at the cut-off, lifting a student who should have got 35% to 40%; that is heaping. Or, the entire distribution could shift up or down, not just those near 40%, which is what the wedge measures. The concern is that the threshold pushing might in aggregate drive

the wedge. Heaping is heaviest at the schools with the lowest grades, which have more students just below 40%: the bottom fifth of schools heap about 5.2 times as often as the top fifth (Appendix K.1). Poorer schools also set easier assessments (Section 4.5). If the schools that push students over 40% were the ones with easy assessments, the wedge would partly measure the pushing. It is the other way round. Low grades are not the same as easy assessments: of two schools whose students do equally well in Grade 8, the one with the lower Grade 7 grades set the harder assessments. Ranked by the wedge, the fifth of Gauteng schools with the hardest assessments heap about 3.4 times as often as the fifth with the easiest in 2022. One reason may be that difficult schools know their assessments are harder than average and so they are reluctant to hold back a student who would have passed at an easier school. The grades bunch at thresholds, but the wedge is not concentrated at these thresholds (Appendix K).

5.3 Smoothing the heap

The heaping is not symmetrical: students' grades 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 the heaping 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 known school effects, and with grades just below 30%, 40% and 50% pushed up to the threshold at a strength set to match the observed heap, is passed to a regression with primary-school and class effects (the simulation records each student's class). The estimator recovers the order as well from the heaped grades as the clean ones: the rank correlation with the truth holds at 0.968, the same 0.968 as with no heaping, and the Grade 7 slope shifts by only 0.003 between heaped and clean grades, even at the strongest heaping simulated. When heaping is made heaviest at the schools with the lowest mean Grade 7 grades, as steeply as in Gauteng's data and more, the rank correlation with the known effects stays within 0.0004 of its value on clean grades and the Grade 7 slope moves by about 0.4 per cent (Appendix K.5). The simulation does not reach the link seen in Gauteng's data between heaping and the wedge itself, where the schools with harder assessments heap more.

On the observed grades, a de-heaping exercise likewise recovers a materially unaltered ranking of schools. Smoothing both grade distributions with a Kleven–Waseem counterfactual density (Kleven and Waseem 2013) (Figure 14) and re-estimating moves the Grade 7 slope by at most 2.1 per cent in any of the six provinces in either cohort, and leaves raw and de-heaped wedges correlated at 0.9973 or above throughout. The counterfactual is a global polynomial: an aggressive smoothing that does not attempt to recover the true distribution but provides a harder target. De-heaping redraws every grade in the two-point windows around 30%, 40% and 50%, 25.0% of the 2022 Grade 7 grades, so it changes more grades than removing the heap would. The downstream regressions show the same covariate gradients under the de-heaped measure as under the baseline (Online Appendix Table 33).

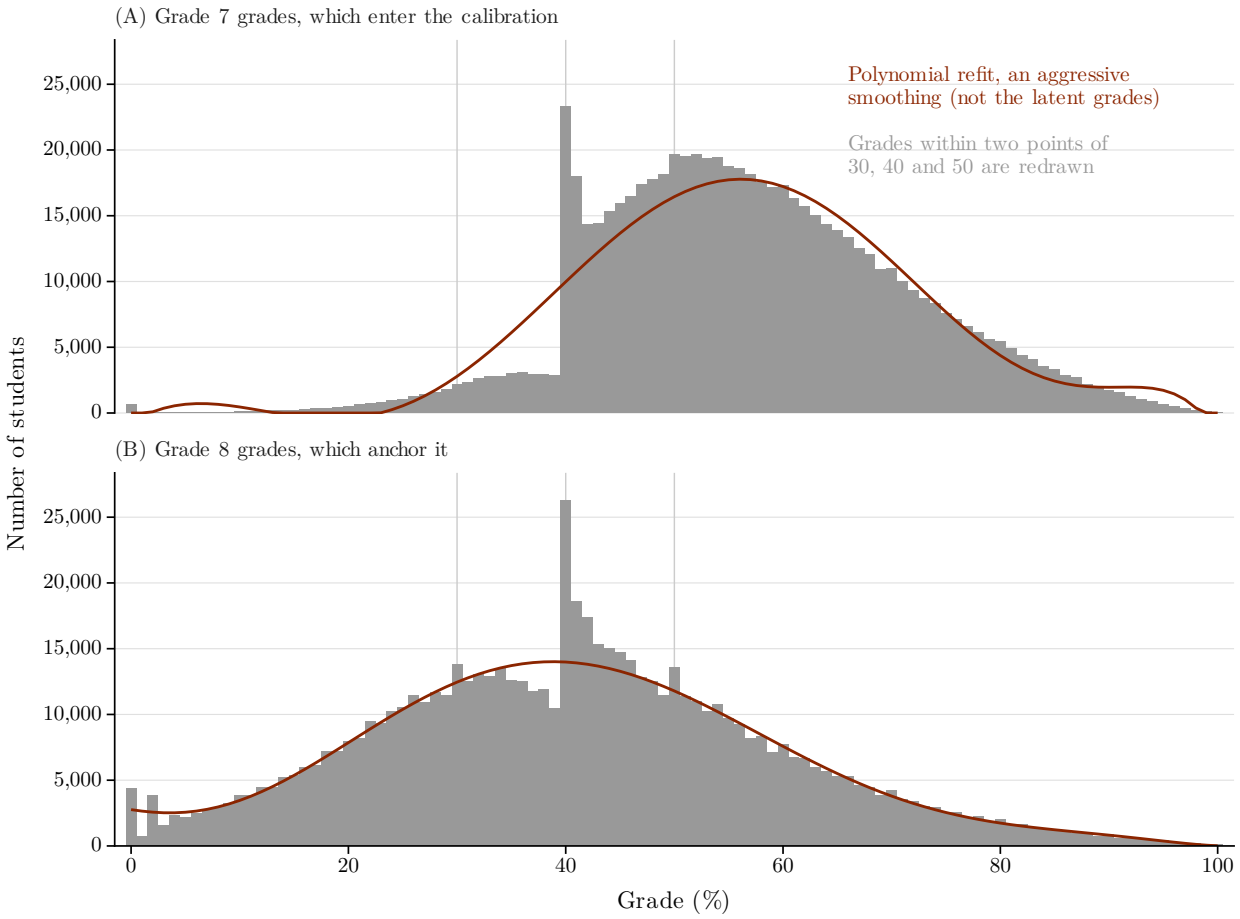

Figure 14: **Grade 7 and 8 grades spike at 40%; an aggressive refit smooths the spike away.**

Notes: Bars are integer-grade frequencies and the line is the polynomial refit, for the Grade 7 grades that enter the calibration (Panel A) and the Grade 8 grades that anchor it (Panel B). The refit is a seventh-order polynomial estimated on the frequencies outside a two-point window around each threshold (30%, 40% and 50%). De-heaping redraws every grade inside those windows from the refit's shape: 25.0% of Grade 7 and 31.1% of Grade 8 grades. Sample: the six provinces, mathematics, 2022 cohort; 665,514 grades in each panel.

6 Conclusion

Many schools set and grade their own assessments. Often, the only comparable result is the secondary school leaving examination, so a school system cannot tell which of its primary schools are behind, or by how much. This paper derives a common scale of learning from the grades schools already award, without a new test. This method requires a large number of recorded school transitions, typical of the primary to secondary transition where demand for primary school comparability might be particularly high.

The Grade 7 grades South Africa's education departments already hold, once calibrated, describe a different school system from the one the raw grades show. Most primary schools change quintile of learning, and support sent to the bottom quintile on raw grades would miss more than three in four of the schools whose students have learned least.

At the same raw grade, students from primaries with easier assessments, whose grades overstate what they have learned, are less likely to sit the NSC five years later. The association is not shown to be causal, so the common scale shows which students need support, not what support would help them. Making easier assessments more accurate may not raise NSC results: some primaries may set easier assessments because they teach at the right level for weaker students (Angrist and Meager 2023), which would also explain those students' lower NSC results. Whether easier assessments reflect lenient grading or teaching at the right level is a question for future research.

The gap between the richest and the poorest primary schools did not narrow between the 2018 and 2022 cohorts, though the raw grades say the gap roughly halved. With calibrated grades, departments can follow such gaps from cohort to cohort, and evaluate a policy by whether the gap between the targeted schools and the rest closed.

A primary school can learn how hard its assessments are compared with other schools', and parents what their child's grades are worth. Each family could be told its own school's wedge; no league table would be published, and the calibrated grades would not be used for secondary-school admission.

Calibrated grades also give researchers learning as an outcome for every student. With linked secondary school leaving results, the calibrated grades make descriptive value added for secondary schools possible: a composite of selection, instruction, and peer effects, for showing how much schools differ and not for accountability. Reclassification events may offer natural experiments on what makes a school's assessments harder or easier: schools changing fee status, poverty quintile, and the entrance of secondary schools in a primary school's feeder zone.

This method travels to any system where a transition mixes a large number of students from different schools, such as lower-secondary transitions in Europe, the move to secondary school in low- and middle-income systems where school-set grades are the only measure of learning that covers every school, and entry to university. Calibration is a route to comparable grades where there is no common examination, not a substitute for one. 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 comparable grades for every student.

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, 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.

Angrist, Noam, and Rachael Meager. 2023. *Implementation Matters: Generalizing Treatment Effects in Education*. EdWorkingPaper No. 23-802. Annenberg Institute at Brown University. https://doi.org/10.26300/ysyd-t435.

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/10.17615/96vs-kw06.

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*.

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 Notice 12, Government Gazette No. 37230, 17 January 2014.

Department of Basic Education. 2016. *2018 School Calendar for Public Schools in South Africa*. Government Notice 1456, 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. *2019 School Calendar for Public Schools*. Government Notice 1154, 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 for Public Schools*. Government Notice 450, Government Gazette No. 44632, 28 May 2021. 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 (ATPs), 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.

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.

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/media

s/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.

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.jdeveco.2014.10.004.

Singh, Abhijeet, and Petter Berg. 2024. "Myths of Official Measurement: Limits to Test-

Based Education Reforms with Weak Governance." *Journal of Public Economics* 239: 105246. https://doi.org/10.1016/j.jpubeco.2024.105246.

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 the WCED Utilises Data*. WCPP 11/4/10.

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.