Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Estimating the causal effect of class size on the achievement of Moroccan students: Evidence from TIMSS 2023

Khalid Boulahris
Mohammed V University
Abdellatif Chatri
Mohammed V University
Said Maizzou
said.maizzou.doc@uhp.ac.ma

Hassan First University

Research Article

Keywords: Class size, Maimonides' rule, Weak instruments, Regression discontinuity, TIMSS 2023, Morocco

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11517111/v1

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

Additional Declarations: No competing interests reported.

Estimating the causal effect of class size on the achievement of Moroccan students: Evidence from TIMSS 2023

Abstract

Background. Administrative caps on class size are the main source of quasi-experimental evidence on class-size effects, but whether a cap generates usable variation depends on the institutional setting. In Morocco, lower secondary classes average 36.5 pupils and a quarter of pupils sit in classes above the legal maximum of 40, yet no study has used the national cap to estimate the causal effect of class size, and whether the cap can identify it is unknown.

Data and methods. We use the TIMSS 2023 Grade 8 sample for Morocco (8,032 pupils, 240 classes, 229 schools) and five subjects: mathematics, physics, biology, chemistry and earth science. Teacher-reported class size is instrumented by the class size predicted from the maximum of 40 pupils and Grade 8 enrollment. A pupil-level design (weighted two-stage least squares with covariates and enrollment controls, and fuzzy regression discontinuity) is combined with a class-level design based on class mean scores. Inference uses the plausible values, school-clustered standard errors and Anderson–Rubin confidence sets robust to weak instruments.

Results. Least squares yields a negative association in every subject, significant in mathematics (-1.76 points per pupil, SE $= 0.47$). In the pupil-level design, the first-stage F statistic falls from 87.4 to 3.9 under a cubic enrollment control, the Anderson–Rubin set is unbounded, and no discontinuity estimate is significant. In the class-level design, the Anderson–Rubin sets exclude zero under linear and piecewise-linear enrollment controls (-4.33 points per pupil in mathematics, set $[-9.8, -0.5]$) but not under a cubic control, and only 12 classes lie near the first three cut-offs. The evidence leans toward a negative effect of class size, but its magnitude is not identified: with a high cap and large cohorts, the rule generates too little local variation.

Keywords: Class size, Maimonides' rule, Weak instruments, Regression discontinuity, TIMSS 2023, Morocco

JEL Classification: I21 , I28 , C26 , C21

1 Introduction

Reducing the number of pupils per class is one of the few schooling reforms whose logic is immediately legible to parents, teachers and legislators alike, and one of the most expensive to implement. Its appeal rests on an intuition about classroom life: with fewer pupils a teacher can attend to each of them more closely, spend less time restoring order and adapt instruction to what pupils need. Four decades of research have nevertheless failed to deliver a context-free estimate of its effect (Ehrenberg, Brewer, Gamoran, & Willms, 2001), and the dispersion of findings reflects both genuine heterogeneity across settings and the difficulty of the underlying identification problem.

That problem is acute in any study based on a single cross-section. Schools operating smaller classes differ systematically from those that do not, and pupils are allocated to classes for reasons related to their prior attainment, so the observed correlation between class size and achievement confounds the effect of the input with the process that assigns it. Exogeneity must therefore come from an identified source of variation rather than from the richness of the covariate set.

The leading such source is the administrative rule capping the number of pupils per class, which generates discontinuities in the relationship between grade enrollment and class size: a cohort of 40 forms one class, a cohort of 41 forms two, and average class size falls abruptly. Angrist and Lavy (1999) used the Israeli maximum of 40 — attributed to Maimonides — to instrument class size and found that smaller classes raised reading and mathematics achievement in the upper primary grades. The design has since been applied to several national systems, with heterogeneous results (Akabayashi & Nakamura, 2014; Gary-Bobo & Mahjoub, 2013; Leuven, Oosterbeek, & Rønning, 2008), and to international assessments (Konstantopoulos & Shen, 2016; Li & Konstantopoulos, 2016; Shen & Konstantopoulos, 2017, 2021). Its re-application to Israel with later cohorts found little evidence of comparable effects (Angrist, Lavy, Leder-Luis, & Shany, 2019).

Morocco has not been among the settings examined, and it differs from all of them in ways that bear directly on the design. Lower secondary classes average 36.5 pupils, a quarter of pupils sit in classes exceeding the legal maximum of 40, and Grade 8 achievement is far below the international centre point — 378 in mathematics and 327 in science against a scale centred on 500 (von Davier, Kennedy, et al., 2024). At the same time the cap is high and grade cohorts are large, which places most schools well away from the discontinuities at which the rule bites hardest. Whether a rule-based instrument retains power under these conditions has not been examined.

Three gaps follow. There is no peer-reviewed estimate of the effect of actual classroom size on achievement among Moroccan pupils; existing Moroccan work cannot separate class size from grade enrollment or from the non-random allocation of pupils to classes (Ibourk, 2013; Liouaeddine, Bijou, & Naji, 2017); and rule-based assessments of class size have been conducted almost exclusively in systems with low caps and small cohorts, leaving the behaviour of the design in the opposite configuration unexamined.

This paper addresses three questions. First, what is the adjusted association between class size and Grade 8 achievement in Morocco, and how much of the raw

relationship is composition? Second, does the Moroccan maximum-class-size rule generate variation in observed class size strong enough to support causal inference, judged by first-stage relevance, the continuity of the enrollment density and the balance of predetermined covariates? Third, what can and cannot be concluded about the causal effect of class size given the answer to the second question?

The contribution is threefold. The study provides the first application of a rule-based design to Moroccan TIMSS 2023 Grade 8 data. It treats the strength of the rule-induced variation as the object of analysis rather than as a maintained assumption, reporting relevance diagnostics for every specification, including those that fail. And it documents an institutional configuration under which the design loses power, which may represent a boundary condition for the portability of rule-based class-size instruments to systems with large cohorts, high caps and imperfect compliance.

The findings are a conditional null, and the conditions carry the result. Least squares yields a negative adjusted association in every subject, significant in mathematics and of a magnitude comparable to the experimental benchmark; that association does not survive instrumentation in the pupil-level design. The first-stage F statistic falls from 87.4 without an enrollment control to 3.9 under the cubic polynomial the design requires, the Anderson–Rubin confidence set is then unbounded, and no discontinuity estimate is significant at any bandwidth between ± 2 and ± 6. The class-level analysis gives a sharper but more fragile picture. Under linear and piecewise-linear enrollment controls the estimates are negative in all five subjects and the Anderson–Rubin sets exclude zero; but the first stage weakens as the control becomes more flexible, the sets become uninformative under the cubic control, and the result cannot be checked locally, because only 12 Moroccan classes lie within five pupils of the first three cut-offs, against 471 in the Israeli data in which the rule was first exploited (Angrist & Lavy, 1999). We do not read this as evidence that class size does not matter in Morocco, nor as an estimate of how much it matters. We read it as evidence that the rule, in this configuration, generates too little local variation for the answer to be independent of how school size is modelled.

Section 2 reviews the literature, Section 3 states the contribution and the research questions, Section 4 describes the data and the measurement of the two key variables, Section 5 sets out the identification strategy and the pupil-level and class-level estimators, and Section 6 the results. Section 7 separates what the study establishes from what it does not, Section 8 states the limitations and Section 9 concludes.

2 Literature review

2.1 Why class size may matter

Class size is best understood as an organizational input that alters classroom processes rather than as a pedagogical treatment in its own right. In the disruption model of Lazear (2001), instruction is a public good within the classroom: teaching is interrupted whenever one pupil is disruptive, so the probability of uninterrupted learning falls as the number of pupils rises. Process studies add that teachers in smaller classes can monitor understanding more often, give individual feedback and organize participation more equitably (Blatchford, Bassett, & Brown, 2011). These mechanisms

do not imply a constant marginal effect: if congestion matters only above a capacity threshold, reductions from very large classes may matter more than reductions within a moderate range, so the effect may be nonlinear and may differ by subject, grade and pupil background.

The same mechanisms make class-size policy contingent on implementation. Smaller classes require more classrooms and more teachers, and a rapid expansion can dilute teacher experience, as the Californian reform illustrated (Jepsen & Rivkin, 2009). In the education-production-function perspective, achievement reflects accumulated family, school, teacher and peer inputs together with prior achievement and unobserved pupil characteristics (Todd & Wolpin, 2003). Selection can therefore bias naive estimates in both directions: better-resourced schools may operate smaller classes, while struggling pupils may be placed in small remedial groups. The central empirical question is which source of variation supports an estimate, and which causal interpretation that variation permits.

2.2 Experimental and quasi-experimental evidence

The strongest experimental evidence comes from Tennessee's Project STAR, which randomly assigned pupils and teachers in the early grades to small or regular classes and found achievement gains, larger for some disadvantaged groups (Krueger, 1999), with some persistence into later outcomes (Krueger & Whitmore, 2001). Its transportability is limited to a large reduction, in early primary grades, in one US state in the 1980s.

Where randomization is unavailable, administrative rules governing class formation have become the main source of quasi-experimental variation. Angrist and Lavy (1999) exploited the maximum of 40 pupils per class applied in Israeli public schools since 1969. Under strict application, class size rises with grade enrollment up to 40 pupils, then falls sharply when a 41st pupil requires a second class, to an average of 20.5, and the pattern repeats at 80, 120 and so on. This saw-tooth function of enrollment serves as an instrument for actual class size, while smooth functions of enrollment absorb any direct effect of school size. The rule explained about half of the variation in class size, and the estimates showed that smaller classes significantly raised reading and mathematics scores in fifth grade and, more modestly, reading scores in fourth grade, but not scores in third grade.

Later rule-based studies produced heterogeneous results: positive effects on sixth-grade language scores in Japan (Akabayashi & Nakamura, 2014), small effects on grade promotion in French junior high schools (Gary-Bobo & Mahjoub, 2013), and effects indistinguishable from zero in Norwegian lower secondary schools (Leuven et al., 2008). Re-examining Israel with recent cohorts, Angrist et al. (2019) found little evidence of comparable effects and documented behavioural responses around the enrollment thresholds. Studies exploiting other variation are similarly dispersed, from no effect in Connecticut (Hoxby, 2000) to long-term educational and labour-market effects in Sweden (Fredriksson, Öckert, & Oosterbeek, 2013). Syntheses reinforce this caution: early meta-analyses reported sizable benefits (Glass & Smith, 1979), whereas design-sensitive reviews found at most a small effect in reading and no robust effect in

mathematics (Filges, Sonne-Schmidt, & Nielsen, 2018), and effects that become negligible outside Project STAR once publication bias is corrected (Opatrný, Havránek, Havránková, & Ščasný, 2025).

2.3 When does a maximum-class-size rule identify a causal effect?

A rule-based design rests on conditions that can largely be tested. First, the rule must exist and bind: schools must actually open an additional class when enrollment crosses a multiple of the maximum. Compliance need not be perfect — the design is fuzzy because space, staffing and parental pressure also shape class formation — but the instrument must be relevant, and the usual benchmark is a first-stage F statistic above 10 (Stock, Wright, & Yogo, 2002). Second, enrollment must not be manipulated around the thresholds. In Chile, a maximum of 45 led many schools to keep enrollment just below its multiples, so that pupils on either side of the cut-offs were no longer comparable (Urquiola & Verhoogen, 2009). Manipulation can be detected by testing the continuity of the enrollment density at the cut-offs (McCrary, 2008) and the balance of predetermined characteristics on either side (Lee & Lemieux, 2010). Third, enrollment may affect achievement through channels other than class size, which requires smooth enrollment controls or narrow windows around the cut-offs, with the attendant trade-off between bias and precision (Imbens & Kalyanaraman, 2012). Finally, the maximum must be fixed by regulation rather than calibrated on the observed data, as was done when rules were unknown (Woessmann, 2005), because calibration may reintroduce the endogenous class formation the design seeks to remove.

What this literature has examined far less is the quantitative question of how much variation a given rule generates in a given enrollment distribution. Relevance is routinely reported as a pass–fail diagnostic after the fact, rather than treated as a property of the institutional configuration that can be anticipated before estimation. Section 5.4 makes that property explicit, and Section 7.2 returns to it.

2.4 Evidence from international large-scale assessments

International large-scale assessments offer common achievement measures across systems and rich questionnaires from pupils, teachers and school leaders. TIMSS is especially relevant because it measures mathematics and science at the fourth and eighth grades in intact sampled classrooms. Its use requires plausible values, sampling weights, replication-based variance estimation and attention to clustering; ignoring these features yields precise-looking but misleading estimates (Caponera & Losito, 2016). Woßmann and West (2006) exploited between-grade variation within schools in TIMSS 1995 and found sizable effects in only two of 11 countries, Greece and Iceland, which they linked to relatively low teacher salaries. That design requires two adjacent grades sampled within schools and cannot be transferred to a single-grade national file.

Several studies have applied rule-based designs to TIMSS and PIRLS, generally finding effects in only a few systems (Konstantopoulos & Shen, 2016; Li & Konstantopoulos, 2016; Shen & Konstantopoulos, 2017). The most closely related study is Shen and Konstantopoulos (2021), who analyzed Grade 8 mathematics and science in

Hungary, Lithuania, Romania and Slovenia across TIMSS 2003, 2007 and 2011. Their country selection is instructive: Norway and Sweden were excluded because they had no national maximum, and Italy because its rule was not followed. The instrument was the class size predicted by the national maximum and grade enrollment, with a cubic function of enrollment and pupil, teacher and school controls; a complementary discontinuity analysis used a binary small-class instrument within windows of three and four pupils around the cut-offs, with covariate-balance tests. Significant achievement effects appeared mainly in Romania in 2003, where one fewer pupil per class raised scores by about 0.05 to 0.09 standard deviations; most other estimates were indistinguishable from zero. The instrument was weak in Hungary, where compliance with the rule was poorer, and first-stage F statistics fell below 10 in most discontinuity samples. The authors also noted that TIMSS records enrollment late in the school year rather than when classes are formed.

2.5 Developing-country evidence and the Moroccan context

In low- and middle-income settings, large classes coexist with teacher shortages, uneven infrastructure and urban–rural gaps. A reduction may relieve genuine congestion, or have little effect if instructional time, teacher attendance or learning materials are more binding. The evidence is accordingly mixed: a positive instrumented class-size coefficient in Bangladesh (Asadullah, 2005), negative effects of larger classes in rural Bolivia (Urquiola, 2006), and little learning gain from class-size reductions over common ranges in India (Datta & Kingdon, 2023). Cross-national evidence nevertheless suggests that class-size effects are larger in developing countries and smaller where teacher quality is higher (Altinok & Kingdon, 2012), which makes Morocco a setting in which effects are plausible rather than presumed.

Moroccan research using international assessments has focused on the broad determinants of achievement. Ibourk (2013) identified pupil characteristics, family environment and school context as the main determinants and emphasized urban–rural and public–private segmentation. Mourji and Abbaia (2013) likewise stressed pupil and household characteristics in TIMSS 2003, and Chatri, Chahbi, and Snihji (2021), analyzing 6,080 Grade 8 pupils in TIMSS 2011, found that individual and family characteristics were the strongest predictors and that school location was the most influential school factor. Bijou and Bennouna (2018) examined public educational expenditure with TIMSS 2015. The closest evidence comes from Liouaeddine et al. (2017), who used a Hausman–Taylor estimator on TIMSS and PIRLS 2011: grade enrollment had no effect on Grade 6 reading ($b = 0.001$, $p = .708$) but a negative effect on Grade 8 mathematics and science (b from -0.050 to -0.030, $p \leq .003$). This measures grade-level congestion rather than classroom size; it is nonetheless directly relevant, because grade enrollment is the running variable of a rule-based design and its direct effect is precisely what smooth enrollment controls must absorb. Lagdiri and Ouazzani Touhami (2026) reported an optimum of 23 to 24 pupils in lower secondary education using PISA 2022, a result that bears on the present setting because Moroccan Grade 8 classes average well above that figure.

Applying a rule-based design to Morocco requires documenting the institutional fact on which it rests: a known maximum for public classes. The Ministry of National

Education bases its theoretical school map on a maximum of 40 pupils per class, which also serves to define overcrowding (Cour des comptes, 2018); a lower maximum of 30 pupils was introduced only for Grades 1 and 2 in 2022–2023 (Médias24, 2022) and does not apply to Grade 8.

2.6 Research gaps

The review points to three gaps. First, there is no peer-reviewed estimate of the effect of actual classroom size on achievement among Moroccan pupils, and none using TIMSS 2023. Second, existing Moroccan estimates cannot separate class size from grade enrollment or from the non-random allocation of pupils and teachers to classes (Ibourk, 2013; Liouaeddine et al., 2017), and no Moroccan study has examined whether an administrative maximum on class size provides exogenous variation. Third, the relevance of rule-based instruments has been assessed almost exclusively in systems with low maxima and small grade cohorts, leaving unexamined how the design behaves where the maximum is high and cohorts are large.

3 The present study

This study estimates the relationship between class size and the mathematics and science achievement of Grade 8 pupils in Morocco, and evaluates whether the national maximum-class-size rule supports a causal interpretation of that relationship. It makes three contributions.

Contribution 1. It provides the first application of a Maimonides-type design to Moroccan data, using the nationally representative TIMSS 2023 Grade 8 sample and five subjects. It combines two levels of analysis on the same data: a pupil-level design, with an instrumental-variable estimator, fixed-window discontinuity estimates and survey weights, and a class-level design based on class means, parsimonious enrollment controls and dummy instruments in narrow enrollment windows. Comparing the two shows how far the conclusions depend on the unit of analysis, the covariate set and, above all, the control for school size. Previous Moroccan work on achievement has relied on multilevel or panel-type estimators that cannot isolate exogenous variation in class size.

Contribution 2. It treats the strength of the rule-induced variation as an object of analysis rather than as a maintained assumption. The relevance of the instrument is reported for every specification, including those in which it fails, and is evaluated with weak-identification-robust inference rather than with conventional standard errors alone. This is a departure from the common practice of reporting a single first-stage statistic for a preferred specification after the specification has been chosen.

Contribution 3. It documents an institutional configuration — a high maximum combined with large grade cohorts, imperfect compliance and substantial dispersion between predicted and actual class size — under which the rule generates little usable variation. This may represent a boundary condition for the portability of rule-based class-size instruments. We do not claim that the design fails in all systems sharing these features: a single country cannot establish that. We claim that the Moroccan case provides evidence consistent with such a boundary, that the relevant quantity is

computable in advance from the enrollment distribution and the maximum, and that researchers contemplating this design in comparable systems have reason to compute it first.

The analysis is organized around three research questions:

RQ1. What is the adjusted association between class size and Grade 8 achievement in mathematics, physics, biology, chemistry and earth science in Morocco, and how much of the raw association is accounted for by pupil, teacher and school composition?
RQ2. Does the Moroccan maximum-class-size rule generate variation in observed class size strong enough to support causal inference, as judged by first-stage relevance under the specification required for identification, the continuity of the enrollment density at the cut-offs, and the balance of predetermined covariates?
RQ3. Given the answer to RQ2, what can and cannot be concluded about the causal effect of class size on achievement in Moroccan lower secondary education?

No directional hypothesis is imposed. One rule governs the reporting throughout: no specification is selected on the basis of its p-value, and every estimate obtained is reported, including those that are not significant and those obtained with a weak instrument.

4 Data

TIMSS has monitored the mathematics and science achievement of fourth- and eighth-grade pupils every four years since 1995; the 2023 assessment is its eighth cycle (von Davier, Fishbein, & Kennedy, 2024). The tests measure curriculum-based competencies on scales comparable across countries and cycles; because each pupil answers only a subset of the item pool, TIMSS produces five plausible values per domain, which carry the measurement uncertainty a single score would ignore (Fishbein, Siegel, Atasever, Cooney, & Gonzalez, 2024); and the assessment is accompanied by questionnaires completed by pupils, teachers and school principals, which supply the class-size, enrollment and background information the design requires.

4.1 Target population and sample design

The Grade 8 target population consists of pupils enrolled in the grade representing eight years of schooling counted from the first year of primary education (Siegel & Foy, 2024). TIMSS samples this population in two stages: schools are selected with probability proportional to size within explicit and implicit strata, school size serving as an implicit stratification variable, and one or more intact classes of the target grade are then selected with equal probability within each sampled school, all pupils in those classes being assessed. The Moroccan Grade 8 file contains 8,745 assessed pupils in 262 sampled classes and 250 schools, with 258 mathematics teacher records. A single class was sampled in 239 schools, two in ten and three in one.

One feature of the Moroccan sample governs the whole analysis. Pupils are taught in intact classes: each sampled pupil belongs to exactly one class, attended for all subjects, and the student–teacher linkage file assigns each pupil one mathematics teacher. Class size is therefore a property of the class rather than of the subject, and

the number reported by the mathematics teacher applies to the five subjects. This differs from systems in which pupils are regrouped by subject and class size varies across subjects (Shen & Konstantopoulos, 2021); in the Moroccan data the teacher covariates are likewise those of the mathematics teacher throughout.

4.2 Analytical sample

The analysis requires grade enrollment from the school questionnaire and class size from the mathematics teacher questionnaire. Neither is imputed: schools with a missing value on either are excluded. Grade 8 enrollment is missing for 10 of the 250 schools and teacher-reported class size for 13. After these restrictions the analytical sample comprises 8,032 pupils in 240 classes and 229 schools. Because the restrictions alter the population to which the results apply, estimates are interpreted as applying to pupils in schools reporting both variables.

4.3 Measurement of grade enrollment

Grade enrollment is both the running variable of the discontinuity design and the input from which the instrument is constructed, so its measurement deserves explicit treatment. The TIMSS school questionnaire records enrollment as of the first day of the month in which testing begins (TIMSS & Center, 2023a), whereas classes are formed at the start of the school year. The quantity that determines how many classes a school opens is therefore not the quantity observed.

Three consequences follow. First, the instrument is constructed from a mismeasured input. Because predicted class size is a deterministic but highly nonlinear function of enrollment, classical measurement error in enrollment does not translate into classical measurement error in the instrument: a school whose enrollment fell from 82 to 78 pupils during the year is recorded on the opposite side of a cut-off from where it stood when classes were formed, and its predicted class size is misassigned by 13 pupils rather than by four. Error of this kind attenuates the first stage and does so disproportionately near the cut-offs, where the design concentrates its identifying variation.

Second, the direction of bias in the second stage cannot be signed from the available information. Attenuation of the first stage inflates the variance of the two-stage estimator and, under weak identification, biases it toward the least squares estimate; but whether measured enrollment errs upward or downward depends on the balance of within-year repetition, transfer and dropout, which the public-use files do not record.

Third, repetition and dropout are not negligible in Moroccan lower secondary education, so the assumption that end-of-year and start-of-year enrollment are close is less comfortable here than in the systems previously studied. The TIMSS public-use files contain no second enrollment measure and no within-year flow variables, so enrollment at class formation cannot be reconstructed. Re-estimation with start-of-year enrollment would require the ministerial school register (*carte scolaire*) matched to the TIMSS school identifiers, which is not publicly available; we identify it as the most valuable extension of the present design. A partial check is nevertheless possible. The sampling frame implies a number of Grade 8 classes for each school (the class

weighting factor multiplied by the number of sampled classes), and in 12 schools the recorded enrollment is either smaller than the number of pupils assessed or implies more than 60 pupils per frame class. Section 6.7 reports the class-level estimates with these schools excluded.

4.4 Measurement of class size

Class size is reported by the mathematics teacher (item 10 of the Grade 8 mathematics teacher questionnaire: "How many students are in this class? Write in the number"; TIMSS and Center 2023b) rather than drawn from an administrative register. Teacher reports are subject to recall error, to ambiguity about pupils who are enrolled but rarely present, and to rounding.

Measurement error in the endogenous regressor has four consequences here. It weakens the first stage, because the rule predicts the administrative class size and not the reported one. It attenuates the least squares coefficient toward zero, so that the adjusted associations of Section 6.2 are if anything conservative. It inflates the standard errors of all estimators. And it widens the observed dispersion between predicted and actual class size, which is the quantity by which compliance with the rule is judged.

This last point matters for interpretation. A school whose reported class size lies well above the rule-predicted value may have failed to open an additional class, or may have opened it and reported the figure imprecisely. The two mechanisms — institutional non-compliance and reporting error — have the same signature in these data and cannot be separated with the available variables. Section 7.2 therefore attributes the observed dispersion to the two jointly, and the manuscript does not claim to measure compliance as such.

4.5 Missing data

Missing values of continuous covariates are replaced by the median and missing values of binary covariates by zero, with indicators for missingness added for the Home Educational Resources scale and for parental education. Missing rates are 5.5% for class size, 4.9% for parental education, 4.7% for school location, 3.6% for Grade 8 enrollment, 3.4% for the Home Educational Resources scale, 2.5% for the percentage of disadvantaged pupils, 2.1% for teacher experience, 1.7% for teacher education, 1.3% for teacher sex and 0.2% for age.

This procedure is common in TIMSS-based class-size research (Shen & Konstantopoulos, 2021), but it is not innocuous and its assumptions should be stated. Median imputation treats the imputed value as observed, so it understates the variance of the covariate, compresses its distribution toward the centre and attenuates its coefficient; it assumes, implicitly, that missingness is unrelated to the outcome conditional on the remaining covariates. Zero-coding a binary covariate is stronger still, since it assigns every non-respondent to one category. The missingness indicators partially relax these assumptions by allowing non-respondents to differ in mean outcome, which is why they are included for the two variables with the least plausible missing-at-random mechanism. Because none of the imputed variables is the treatment or the instrument —

both of which are left missing and their observations dropped — the procedure affects the estimates only through the covariate adjustment.

Section 6.8 reports a complete-case analysis as a robustness check. Multiple imputation would in principle be preferable to either approach, since it propagates imputation uncertainty into the standard errors, but it interacts awkwardly with the five plausible values, which are themselves an imputation: a correct implementation requires imputations nested within each plausible value and a nested combination of the two sources of variance (Schafer, 1999). We do not implement it. Because neither the treatment nor the instrument is imputed, and because the complete-case estimates in Section 6.8 are close to the main ones, there is little reason to expect it to alter the conclusions; it remains a natural refinement of the covariate adjustment.

4.6 School sector

The maximum of 40 pupils is a norm of the public system, and private schools are not bound by it in the same way, so compliance should in principle be assessed within the public sector. The TIMSS 2023 public-use files for Morocco contain no indicator of school sector, and no variable in the school questionnaire functions as a reliable proxy: school location, enrollment and the percentage of disadvantaged pupils are all correlated with sector but none identifies it. A sector-specific analysis therefore cannot be conducted with the available variables. The consequence is that the sample pools schools subject to the rule with schools that may not be, which attenuates the first stage by an amount that cannot be quantified here, and that the estimates apply to the Moroccan Grade 8 population as a whole rather than to its public component. Section 8 returns to the implications for external validity.

4.7 Class-level sample

The class-level analysis takes the class as the unit of observation, with the class mean score as the outcome. It is estimated on the 218 schools of the analytical sample in which a single class was sampled, which together contain 7,393 pupils. In these schools the class is the school's only sampled Grade 8 class, so the class mean, the teacher-reported class size and the school-level instrument refer to the same unit, and each observation is an independent school. The 11 schools of the analytical sample with two or three sampled classes are excluded from this part of the analysis; their inclusion would add little information and require a correction of the standard errors for within-school correlation. Class means are computed separately for each plausible value and the five resulting estimates are combined by the multiple-imputation rules. The only school characteristic used in the class-level models is the percentage of economically disadvantaged pupils (PD), which TIMSS records in four categories (0–10%, 11–25%, 26–50% and more than 50%) rather than as a continuous percentage; it is entered as a score from 1 to 4. Appendix Table A1 reports descriptive statistics for this sample.

5 Identification and estimation

5.1 The institutional rule

Moroccan schools operate under an administrative maximum of 40 pupils per class. The Ministry of National Education uses this threshold to build its theoretical school map and to define overcrowded classes (Cour des comptes, 2018). The maximum is taken from regulation and is not calibrated on the observed class-size distribution, which avoids reintroducing the endogenous class formation the design seeks to remove (Woessmann, 2005).

5.2 Predicted class size

If a school enrolls e_j pupils in Grade 8 and no class may exceed R pupils, the rule requires $\text{int}((e_j - 1)/R) + 1$ classes, and the average class size it predicts is

$$f_j = \frac{e_j}{\text{int}\left(\dfrac{e_j - 1}{R}\right) + 1} \tag{1}$$

where $\text{int}(\cdot)$ is the largest integer less than or equal to its argument. With $R = 40$, a school enrolling 40 pupils forms one class of 40; a school enrolling 41 must form two and the predicted average falls to 20.5; the predicted value rises again to 40 at 80 pupils and falls to 27 at 81. The instrument takes the same value for all Grade 8 classes of a school.

5.3 Enrollment, predicted class size and actual class size

Equation (1) makes predicted class size a deterministic saw-tooth function of enrollment: strictly increasing within each segment between consecutive multiples of R, and discontinuous at each multiple. Two distinct relationships must therefore be kept apart.

The first is between enrollment and predicted class size. It is mechanical and exact by construction; it is not an empirical finding and it carries no identifying content on its own.

The second is between predicted and actual class size. This is an empirical relationship and it is what the first stage measures. In the Moroccan sample the correlation between the two is 0.43 and 53% of schools lie within three pupils of the predicted value, with alignment close among schools forming one or two classes and deteriorating as schools grow (Figure 1). This relationship is the one that must be strong for the instrument to be relevant — and, crucially, it must be strong *after* conditioning on enrollment, which is a different requirement from being strong unconditionally.

5.4 Relevance and the exclusion restriction

The two conditions an instrument must satisfy pull in opposite directions in this design, and separating them is the central methodological point of the paper.

Relevance requires that the instrument predict the endogenous regressor conditional on the covariates included in the second stage. **The exclusion restriction** requires that the instrument affect the outcome only through the endogenous regressor. Because f_j is a deterministic function of e_j, the exclusion restriction cannot hold unconditionally: grade enrollment plausibly affects achievement through channels other than class size — school organisation, resource allocation, peer composition, the composition of the teaching staff — and the Moroccan evidence suggests that it does, since Liouaeddine et al. (2017) report a direct negative effect of Grade 8 enrollment on mathematics and science achievement in Moroccan TIMSS data. The design therefore requires smooth functions of enrollment in the covariate vector, whose role is to absorb any direct effect of school size and leave only the discontinuities of the rule as the source of identification (Angrist & Lavy, 1999; Lee & Lemieux, 2010).

The cubic polynomial is retained as the reference specification. Its justification is substantive rather than statistical: a polynomial flexible enough to approximate a smooth, possibly non-monotonic relationship between school size and achievement over a range from 17 to 794 pupils, while remaining a global smooth function and therefore unable to reproduce the saw-tooth itself. The sequence of enrollment controls — none, linear, cubic — is reported in full so that the consequence of each choice is visible.

The tension is then immediate. Every smooth function of enrollment added to the covariate vector strengthens the case for the exclusion restriction and weakens the first stage, because the instrument is itself a function of enrollment and only its non-smooth component survives the conditioning. A high first-stage F obtained without the enrollment control is therefore not evidence of a valid instrument: it measures the joint predictive power of the rule and of school size, which is precisely the mixture the design is meant to separate. A strong unconditional relationship between the rule and class size does not establish a valid instrument for the causal effect of class size, and reporting relevance without stating the conditioning set is uninformative.

5.5 Estimation

Least squares. For pupil i in school j,

$$Y_{ij} = \alpha_0 + \alpha_1\, ClassSize_{ij} + ST_{ij}A_2 + TCH_{ij}A_3 + SCH_jA_4 + u_{ij} \tag{2}$$

where ST contains pupil sex, age, home educational resources and parental education; TCH teacher sex, experience and education; and SCH the percentage of economically disadvantaged pupils, school location and, in the third specification, Grade 8 enrollment in linear, quadratic and cubic form. The coefficient α_1 is an *adjusted association*: pupils are not assigned to classes at random, and nothing in Equation (2) licenses a causal reading. It is reported because the sequence of specifications shows how much of the raw relationship is composition, and because it is the benchmark against which the instrumented estimates are read.

Two-stage least squares. The first stage regresses class size on the instrument and the covariates,

$$ClassSize_{ij} = \pi_0 + \pi_1 f_j + ST_{ij}\Pi_2 + TCH_{ij}\Pi_3 + SCH_j\Pi_4 + \varepsilon_{ij} \tag{3}$$

and the second stage regresses achievement on the fitted values,

$$Y_{ij} = \beta_0 + \beta_1 FV_{ij} + ST_{ij}B_2 + TCH_{ij}B_3 + SCH_jB_4 + e_{ij} \tag{4}$$

where β_1 is the parameter of interest and a negative value indicates that achievement rises as class size falls.

Weak-identification inference. Relevance is assessed by the F statistic on π_1 in Equation (3), computed with standard errors clustered at the school level, and compared with the conventional bound of 10 (Stock et al., 2002). With a single instrument and a single endogenous regressor, this cluster-robust statistic coincides with the Kleibergen–Paap rank Wald statistic and with the effective first-stage F of Montiel Olea and Pflueger, whose stricter benchmark of about 23 is also reported. Because conventional two-stage inference is unreliable when the first stage is weak, three further procedures are applied to the mathematics outcome: the limited-information maximum likelihood and Fuller(1) estimators, which have better finite-sample properties than two-stage least squares under weak identification; and Anderson–Rubin confidence sets obtained by grid inversion, which are valid regardless of instrument strength. The Anderson–Rubin set is constructed by regressing $Y_{ij} - \beta_0 ClassSize_{ij}$ on the instrument and the covariates for each candidate β_0 on a grid, retaining the values at which the coefficient on the instrument is not rejected at the 5% level, and combining the five plausible values before inversion.

5.6 The discontinuity design as a local diagnostic

The rule also generates a discontinuity in class size: when enrollment passes a multiple of R by one pupil an additional class must be opened, so average class size is smaller just to the right of each cut-off than just to the left. Around these cut-offs the design resembles a local randomised experiment, and because factors other than the rule also influence class formation it is fuzzy rather than sharp, estimated with the two-stage approach of Equations (3) and (4) with a binary instrument.

Its role in this paper is diagnostic rather than definitive, for two reasons. First, restricting the sample to narrow windows around the cut-offs reduces it to a few dozen schools, so the design trades the bias from enrollment controls for a loss of precision severe enough to make the resulting estimates uninformative about magnitudes. Second, the discontinuity sample permits tests that the full sample does not — covariate balance, density continuity, placebo cut-offs — and those tests, rather than the point estimates, are what the exercise contributes. The discontinuity results are therefore read as evidence about whether the design is credible, not as a second estimate of the effect.

Two implementations are reported. The first uses fixed windows: for a bandwidth h the sample includes schools whose enrollment lies in the h values immediately to the left of a cut-off, from $kR - h + 1$ to kR, and in the h values immediately to its right, with a binary instrument equal to 1 on the right. A stricter variant retains only the segments in which a single threshold on predicted class size separates the two sides of every cut-off; with $R = 40$ and $h = 4$ this selects the first ten cut-offs, from 40 to 400 pupils. Restricting the windows to these ten cut-offs, in the single-class schools of

Section 4.7 and with the covariates available without the student questionnaire (pupil sex and age, the teacher covariates, PD and the cubic in enrollment), leaves the first stage equally degenerate ($F = 0.04$ at ± 4 and 0.23 at ± 3). Bandwidths from ± 2 to ± 6 are reported, which brackets the narrow windows of three to five pupils common in this literature (Angrist & Lavy, 1999; Shen & Konstantopoulos, 2021). The second is a local polynomial estimator with data-driven bandwidth selection, reported for polynomial degrees one and two and for triangular and uniform kernels. Neither specification is selected on its p-value: the fixed-window results at ± 3 and ± 4 are designated in advance as the reference, because they are the windows in which covariate balance is assessed and the only ones in which the first stage is not numerically degenerate.

Validity tests. Covariate balance is assessed by

$$Covariate_i = \delta_0 + \delta_1 D_i + \upsilon_i \tag{5}$$

where D_i equals 1 if enrollment lies to the right of a cut-off, with enrollment centred on each cut-off and the data pooled across cut-offs, and standard errors clustered at the school level. Manipulation of the running variable is assessed by the local polynomial density test of Cattaneo, Jansson and Ma applied to school-level centred enrollment, and by binomial tests of the counts on either side of the cut-offs. Finally, the first stage is re-estimated at false cut-offs displaced by 10, 20 and 30 pupils from the true ones: a rule that operates as claimed should produce no jump in class size where no rule applies.

5.7 Survey design, weights and variance

Full-sample models are estimated with the overall student weight (TOTWGT), which projects estimates to the national population of Moroccan Grade 8 pupils. Estimates and standard errors are computed separately for each of the five plausible values and combined by the multiple-imputation rules, with the usual $(1+1/m)$ correction for $m = 5$ (Schafer, 1999). The discontinuity estimates are unweighted, because they are local comparisons around the cut-offs rather than projections to the national population.

Variance estimation requires explicit discussion, since it is a recurrent weakness of the applied literature (Caponera & Losito, 2016). The main results use standard errors that are heteroscedasticity-robust and clustered at the school level. This captures the dominant feature of the TIMSS design — the school is the primary sampling unit — but it is not the jackknife repeated replication (JRR) procedure that the IEA recommends, which uses the 125 sampling zones and the replicate indicators supplied with the data and thereby reflects the stratification as well as the clustering. To establish that the choice is not consequential, the mathematics results were re-estimated with JRR over the 250 replicates. The jackknife and clustered standard errors are close: 0.476 against 0.474 for the least squares estimate, 4.219 against 3.870 for the two-stage estimate, and the first-stage F is 4.07 against 3.94. The jackknife standard error on the instrumented estimate is thus about 9% larger, which widens the confidence interval without altering any conclusion; it is reported alongside the clustered standard error in Table 5. Stratification is reflected in the JRR figures through the zone structure; explicit strata are not separately available in the public-use file.

5.8 Class-level estimation

The class-level analysis complements the pupil-level design with a more parsimonious specification. For class c in school s, the structural equation and the first stage are

$$\bar{Y}_c = \beta_0 + \beta_1 n_c + \gamma PD_s + g(e_s) + u_c, \qquad n_c = \pi_0 + \pi_1 f_s + \delta PD_s + h(e_s) + \xi_c, \quad (6)$$

where $\bar{Y}_c$ is the class mean score, n_c teacher-reported class size, f_s the predicted class size of Equation (1), PD_s the percentage of disadvantaged pupils, and $g(\cdot)$ and $h(\cdot)$ functions of enrollment. Least squares, the first stage, the reduced form and two-stage least squares are estimated on the full class-level sample under four specifications: (1) PD only; (2) PD and enrollment; (3) PD, enrollment and enrollment squared divided by 100; and (4) a continuous piecewise linear trend in enrollment with no other covariate. The trend has slope $1/k$ on the kth enrollment segment, so that variation around it is generated only by the jumps of the rule:

$$T(e) = \frac{e}{k} + \sum_{j=2}^{k} \frac{R}{j}, \qquad k = \operatorname{int}\left(\frac{e-1}{R}\right) + 1. \quad (7)$$

This trend was proposed for enrollments up to 160 pupils (Angrist & Lavy, 1999); Moroccan cohorts are larger and the formula is applied over the whole enrollment range.

Specifications (5) and (6) re-estimate (1) and (2) in a ± 5 discontinuity sample that retains only schools close to the first three cut-offs, where the prescribed reduction in class size is largest, with enrollment in $[36, 45]$, $[76, 85]$ or $[116, 125]$. A dummy-instrument estimator is also applied in this sample and in a ± 3 sample, with enrollment in $[38, 43]$, $[78, 83]$ or $[118, 123]$. Class size is instrumented by three dummies $z_{js} = 1[\, e_s > 40j \,]$ restricted to segment j, which mark the upper half of each segment, and the only control for enrollment consists of dummies for the first two segments. Because predicted class size is below 32 whenever one of the z_{js} equals one and above 32 otherwise, this instrument set is the dummy $1[f_s < 32]$ interacted with segment, and the estimator is a weighted combination of within-segment Wald estimates.

The class-level models are unweighted, the outcome is the class mean of each plausible value, the five estimates are combined by the multiple-imputation rules, and standard errors are heteroscedasticity-robust, which with one class per school is equivalent to clustering by school. The pupil-level and class-level designs therefore differ in four respects — the unit (pupil or class), the covariates (pupil, teacher and school characteristics, or PD alone), the weighting, and the enrollment control (cubic, or linear, quadratic and piecewise linear) — and together they bracket the range of specifications used in the literature. To separate the role of the enrollment control from the other three differences, a bridge specification adds the cubic polynomial of the pupil-level design to the class-level model with PD. For each class-level specification, Anderson–Rubin 95% confidence sets are obtained by exact inversion of the test: for a candidate value b_0, the reduced-form coefficient of f_s in a regression of $\bar{Y}_c - b_0 n_c$ on f_s and the controls is combined across plausible values by the multiple-imputation rules, and b_0 is retained if this coefficient is not significantly different from zero at the 5%

level. Because class size does not vary across plausible values, the acceptance region is the solution of a quadratic inequality in b_0 and can be a bounded interval, the union of two rays or the whole real line.

6 Results

6.1 Descriptive statistics

The analytical sample comprises 8,032 pupils in 240 classes and 229 schools, and is identical across the five subjects because each pupil belongs to a single class. Table 1 reports descriptive statistics. The average class holds 36.45 pupils with a standard deviation of 6.74; the smallest has 10 and the largest 55, and 25.1% of pupils are in classes above the maximum of 40. Grade 8 enrollment has a median of 245 pupils and ranges from 17 to 794. Mean achievement ranges from 317.6 in biology to 377.0 in mathematics.

Table 1 Descriptive statistics

Variable	Mean	SD	Minimum	Maximum
Class size (all subjects)	36.45	6.74	10	55
Grade 8 enrollment (school level)	264	155	17	794
Mathematics score	377.0	74.5	179.4	650.0
Physics score	323.3	88.1	58.9	631.7
Biology score	317.6	90.0	74.3	637.1
Chemistry score	332.9	86.9	76.3	633.0
Earth science score	326.4	93.5	53.1	652.6

N = 8,032 pupils in 240 classes and 229 schools. Weighted statistics (TOTWGT); achievement averaged across the five plausible values. Enrollment statistics are unweighted at the school level; its median is 245.

6.2 Adjusted associations

Table 2 reports the least squares estimates, with covariates added in sequence. The coefficient is negative in every subject and every column and attenuates as controls are added: in mathematics it falls from −2.65 without covariates to −1.70 with pupil, teacher and school characteristics and to −1.76 once enrollment is controlled. The same pattern holds in the four science subjects, where the coefficient roughly halves between columns (1) and (3). With the full set of controls the mathematics coefficient remains significant (−1.76, SE = 0.47), earth science is borderline (p = .065), and physics, biology and chemistry are not significant.

Two readings of these numbers are available and only one is warranted. The attenuation between columns shows that part of the raw gap reflects composition, consistent with better-resourced schools operating smaller classes while weaker pupils are grouped into small remedial classes (Todd & Wolpin, 2003). The residual coefficient is an adjusted association, not an effect. Expressed in standard-deviation units

it corresponds to −0.024 SD per pupil, so that a reduction of eight pupils — the contrast of Project STAR — would correspond to about 0.19 SD, close to the 0.20 SD experimental benchmark (Krueger, 1999). The comparison is useful as a magnitude check and as motivation for seeking an identification strategy; it is not evidence that a reduction of eight pupils would deliver 0.19 SD in Morocco, and the remainder of the paper is addressed to why these data cannot answer that question.

Table 2 Least squares estimates: adjusted associations between class size and achievement

Subject	(1) No covariates	(2) + pupil, teacher, school	(3) + enrollment (cubic)
Mathematics	−2.648*	−1.703*	−1.759*
	(0.593)	(0.432)	(0.474)
Physics	−2.592*	−1.289*	−1.088
	(0.685)	(0.654)	(0.767)
Biology	−2.855*	−1.487	−1.235
	(0.800)	(0.808)	(0.953)
Chemistry	−2.387*	−1.374*	−1.145
	(0.626)	(0.646)	(0.743)
Earth science	−2.780*	−1.573*	−1.468
	(0.676)	(0.653)	(0.797)

$N = 8{,}032$. Standard errors in parentheses, robust and clustered at the school level, combined across the five plausible values. Weighted estimates (TOTWGT). Coefficients are adjusted associations and have no causal interpretation. $^{*}p < .05$.

6.3 First stage and instrument relevance

Figure 1 plots average class size against Grade 8 enrollment with the rule-predicted schedule of Equation (1) superimposed. Compliance is close among schools forming one or two classes and weakens as schools grow: from five classes onward the points lie systematically above the schedule. Table 3 reports the first stage under the three enrollment controls.

Table 3 First-stage regressions of class size on rule-predicted class size

Enrollment control	π_1	SE	F	Identification
None	0.846	0.090	87.37	Strong, but exclusion restriction violated
Linear	0.590	0.116	25.98	Strong
Cubic	0.286	0.144	3.94	Weak

$N = 8{,}032$. Weighted (TOTWGT), standard errors clustered at the school level. With one instrument and one endogenous regressor the cluster-robust F coincides with the Kleibergen–Paap rank Wald statistic and with the Montiel Olea–Pflueger effective F. Benchmarks: 10 (Stock et al., 2002); 23.1 for 10% relative bias under the effective-F criterion. The same first stage applies to all five subjects.

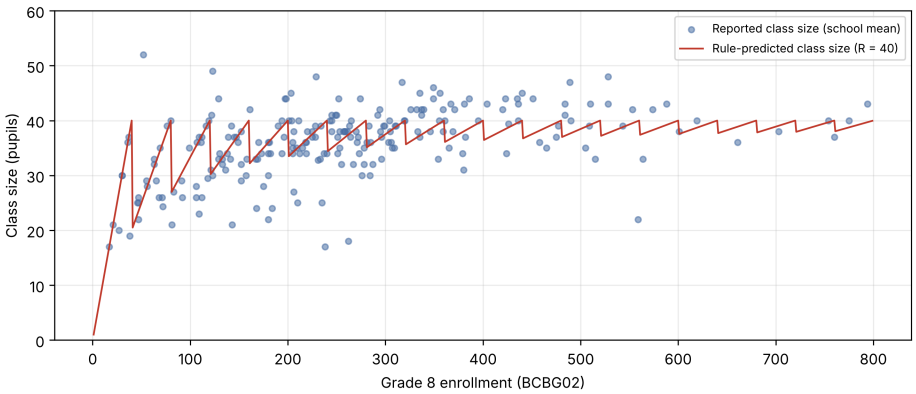


Fig. 1 Actual and rule-predicted class size against Grade 8 enrollment. Points are school-level average teacher-reported class size; the line is the schedule predicted by Equation (1) with $R = 40$.

The pattern in Table 3 is the central empirical finding of the paper. The instrument is a strong predictor of class size when enrollment is left uncontrolled, and remains strong under a linear control, but loses almost two-thirds of its coefficient and nearly all of its explanatory power once the cubic polynomial is included. This is the tension set out in Section 5.4 made quantitative: the specification that makes the exclusion restriction defensible is the specification under which the instrument ceases to be relevant. The first column is reported for completeness and is not a candidate specification.

6.4 Two-stage estimates and weak-identification inference

Table 4 reports the two-stage estimates for the five subjects under the three enrollment controls. Without an enrollment control the mathematics estimate is -1.92 (SE $=$ 0.91) and significant; under a linear control all five estimates are negative, from -1.89 in earth science to -2.58 in mathematics, and none is significant; under the cubic polynomial the estimates are smaller and less consistent in sign and none is significant. The only significant estimate in the table is thus obtained under the one specification the identification argument excludes.

Conventional two-stage inference is not reliable at a first-stage F of 3.9, for three reasons. The two-stage estimator is biased toward the least squares estimate in finite samples, and the bias grows as relevance falls, so the negative sign of the column (6) estimate carries little information — it is the sign the least squares estimate already has. The sampling distribution of the estimator is not well approximated by a normal distribution, so the reported standard error understates the true uncertainty and the associated p-value is not interpretable at face value. And failing to reject the null is not evidence that the parameter is zero: under weak identification the test has little power against any alternative.

Table 5 therefore reports weak-identification-robust results for mathematics. Three features deserve comment. First, the limited-information maximum likelihood estimator coincides exactly with two-stage least squares, and Fuller(1) differs only in the third

Table 4 Two-stage least squares estimates by enrollment control

Subject	(4) No enrollment control	(5) Linear	(6) Cubic
Mathematics	−1.920*	−2.576	−2.858
	(0.908)	(1.593)	(3.870)
Physics	−1.665	−2.115	−0.155
	(1.206)	(1.993)	(5.271)
Biology	−1.937	−2.362	−0.331
	(1.442)	(2.391)	(6.211)
Chemistry	−1.837	−2.161	−0.911
	(1.113)	(1.860)	(4.765)
Earth science	−1.588	−1.889	0.892
	(1.094)	(1.802)	(5.297)
First-stage F	87.37	25.98	3.94

N = 8,032. Standard errors in parentheses, robust and clustered at the school level, combined across the five plausible values. Weighted (TOTWGT). Column (4) does not satisfy the identification requirement of Section 5.4 and is reported for completeness only. *$p < .05$.

decimal: with a single instrument the model is just identified, so these estimators offer no protection here, a point worth stating because they are routinely recommended as a remedy for weak instruments without the qualification that they help only in the over-identified case. Second, the Anderson–Rubin confidence set, which is valid whatever the strength of the instrument, is bounded under the linear control at [−5.75, 0.75] — that is, [−0.077, 0.010] standard deviations per pupil — and is unbounded under the cubic specification. An unbounded confidence set is the formally correct expression of what the data support: every value of the parameter is consistent with the evidence once the exclusion restriction is taken seriously. Third, the jackknife standard error is about 9% larger than the clustered one, which does not change this reading.

6.5 Discontinuity estimates

Figure 2 shows the distribution of Grade 8 enrollment with the cut-offs marked, and Figure 3 shows class size and mathematics achievement against enrollment centred on the nearest multiple of 40. Two features are visible. Schools are present on both sides of every cut-off, unlike the Chilean case in which they clustered to the left (Urquiola & Verhoogen, 2009). And neither panel of Figure 3 displays a visually compelling discontinuity at zero: the drop in class size that the rule prescribes is not apparent in the binned means, which anticipates the first-stage results below.

Table 6 reports the fixed-window estimates at bandwidths from ±2 to ±6, with the first-stage coefficient and F statistic, the second-stage coefficient, its standard error, confidence interval and p-value, and the sample size at each bandwidth. The table is reported in full and invites a blunt reading: outside the ±3 and ±4 windows the estimator is numerically degenerate. At ±2, ±5 and ±6 the first-stage coefficient is near zero or of the wrong sign and the F statistic falls below 0.1, so the second-stage estimates — 13.5, −52.1 and 33.2, with standard errors of 75, 907 and 291 — are ratios

Table 5 Weak-identification diagnostics and robust inference, mathematics

	No enrollment control	Linear	Cubic
First-stage π_1	0.846	0.590	0.286
First-stage SE	(0.090)	(0.116)	(0.144)
F (Kleibergen–Paap rk; effective F)	87.37	25.98	3.94
Exceeds 10 / exceeds 23.1	yes / yes	yes / yes	no / no
2SLS	−1.920	−2.576	−2.858
	(0.908)	(1.593)	(3.870)
LIML	−1.920	−2.576	−2.858
	(0.908)	(1.593)	(3.870)
Fuller(1)	−1.920	−2.575	−2.852
	(0.908)	(1.591)	(3.847)
2SLS, jackknife SE	—	—	(4.219)
Anderson–Rubin 95% set	—	[−5.75, 0.75]	unbounded
in SD units	—	[−0.077, 0.010]	unbounded

N = 8,032. The model is just identified, so LIML coincides with 2SLS by construction. Anderson–Rubin sets obtained by grid inversion over $[-40, 40]$ in steps of 0.25, combining the five plausible values; "unbounded" indicates that the accepted region reaches the limits of the grid. The jackknife standard error uses the 125 TIMSS sampling zones and 250 replicates. Effect sizes use the weighted mathematics standard deviation of 74.5.

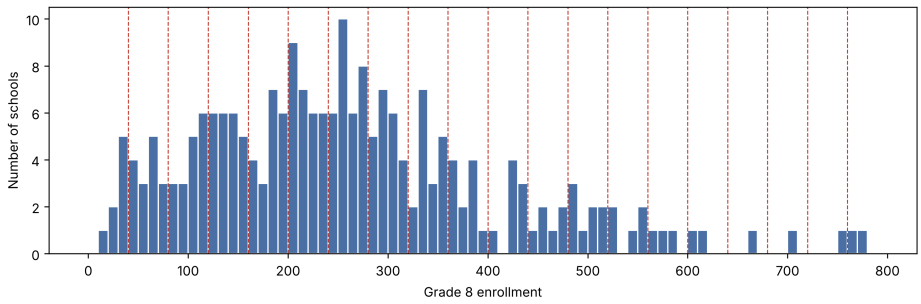


Fig. 2 Distribution of Grade 8 enrollment and cut-offs (multiples of 40).

of two quantities that are both indistinguishable from zero. They are reported because suppressing them would misrepresent the behaviour of the design, not because they contain information about the parameter.

The five-subject estimates at the two reference bandwidths are reported in Table 7. At ±3 every estimate is negative, between −0.07 and −0.15 standard deviations per pupil; at ±4 they are smaller and less consistent in sign; none is significant, the smallest p-value being .17.

Table 8 reports the local polynomial estimates with data-driven bandwidth selection, for polynomial degrees one and two and for triangular and uniform kernels. All four point estimates are negative, from −0.82 to −8.81, none is significant, and the robust confidence intervals span zero widely in every case. The selected bandwidths

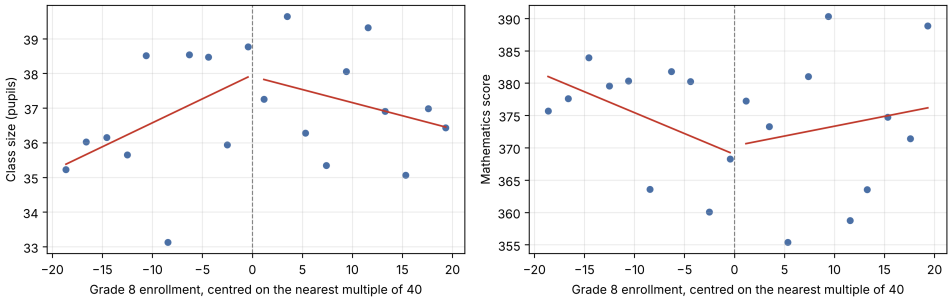


Fig. 3 Class size (left) and mathematics achievement (right) against Grade 8 enrollment centred on the nearest multiple of 40. Points are means within two-pupil bins; lines are separate linear fits on either side of the cut-off.

Table 6 Fixed-window discontinuity estimates by bandwidth, mathematics

h	N	Schools	π_1	F	Estimate	SE	95% CI	p
±2	665	18	−0.584	0.06	13.52	75.48	[−134.4, 161.5]	.858
±3	1,122	31	2.297	1.51	−5.43	4.65	[−14.6, 3.7]	.243
±4	1,568	43	1.423	0.71	0.78	8.40	[−15.7, 17.2]	.926
±5	2,150	59	0.082	0.00	−52.07	907.00	[−1829.8, 1725.6]	.954
±6	2,484	69	−0.157	0.02	33.15	290.63	[−536.5, 602.8]	.909

Unweighted, binary instrument, covariates as in Equation (4), standard errors clustered at the school level and combined across the five plausible values. The ±3 and ±4 windows are the designated reference specifications, chosen before estimation because they are the windows in which covariate balance is assessed; the remaining rows are reported for completeness.

range from 3.5 to 7.3 pupils, which brackets the fixed windows used above and indicates that the fixed-window choice is not driving the result. Figure 4 displays the fixed-window estimates and their confidence intervals across bandwidths, and makes the instability visible at a glance.

6.6 Tests of the discontinuity assumptions

Three sets of tests assess whether the discontinuity design is credible, independently of the precision of its estimates.

Covariate balance. Table 9 reports the estimated discontinuity in each of the eleven covariates at both reference bandwidths. At ±3 no covariate differs significantly across the cut-offs, which supports the local randomisation interpretation. At ±4 four covariates are unbalanced — pupil sex ($p = .029$), pupil age ($p = .020$), parental education ($p = .026$) and teacher sex ($p = .038$) — so the wider window gains sample size at the cost of the property that makes the design attractive. This is the methodological, rather than statistical, ground on which ±3 is preferred to ±4.

Density of the running variable. The local polynomial density test applied to school-level centred enrollment does not reject continuity at the cut-offs ($T = -0.70$,

Table 7 Discontinuity estimates at the reference bandwidths, five subjects

	Bandwidth ±3		Bandwidth ±4		Effect size
Subject	Estimate	SE	Estimate	SE	(±3)
Mathematics	−5.434	4.653	0.782	8.398	−0.071
Physics	−11.255	8.258	−6.972	10.423	−0.126
Biology	−13.374	10.282	−8.451	13.107	−0.145
Chemistry	−10.896	8.401	−7.345	10.742	−0.125
Earth science	−11.527	8.419	−7.106	10.117	−0.122

$N = 1{,}122$ (bandwidth 3) and 1,568 (bandwidth 4). Unweighted; standard errors robust, clustered at the school level and combined across the five plausible values. Effect size = estimate divided by the standard deviation of the outcome in the discontinuity sample. No estimate is significant at $p < .05$. First-stage F: 1.51 at ±3 and 0.71 at ±4.

Table 8 Local polynomial discontinuity estimates, mathematics

Degree	Kernel	Bandwidth	N left	N right	Estimate (SE)	Robust 95% CI
1	Triangular	7.24	1,499	1,527	−0.82 (6.05)	[−11.6, 12.1]
1	Uniform	3.52	629	722	−8.81 (9.33)	[−26.9, 9.7]
2	Triangular	7.30	1,499	1,527	−5.28 (8.08)	[−22.7, 9.0]
2	Uniform	5.77	1,133	1,222	−8.77 (11.84)	[−33.1, 13.4]

Fuzzy specification with class size as the endogenous regressor, enrollment centred on the nearest multiple of 40, covariates as in Equation (4), clustered at the school level, mean-squared-error-optimal bandwidth selection and robust bias-corrected inference. Robust p-values: .964, .358, .395 and .406 respectively.

$p = .483$). Binomial tests of the school counts on either side agree: 18 against 13 within ±3 ($p = .473$), 26 against 19 within ±4 ($p = .371$) and 32 against 25 within ±5 ($p = .427$). There is thus no evidence that Moroccan schools hold enrollment below the multiples of the maximum, in contrast to the Chilean case. This is the one identification condition that the Moroccan data satisfy cleanly.

Placebo cut-offs. Re-estimating the first stage at false cut-offs displaced from the true ones yields no jump in class size at a displacement of 10 pupils (0.51, SE = 1.88, $p = .785$) or 20 pupils (0.80, SE = 1.21, $p = .509$), but a significant jump of −3.51 pupils at a displacement of 30 ($p = .019$). Two of three placebos therefore pass and one fails. The failure is a genuine concern and is reported as such: it is consistent with the saw-tooth structure of the enrollment distribution producing spurious jumps where the rule does not apply, which would indicate that the identifying variation is contaminated by features of school size unrelated to the cap. With three tests the evidence is limited, and the result neither validates nor invalidates the design on its own; it does weigh against treating the discontinuity estimates as causal.

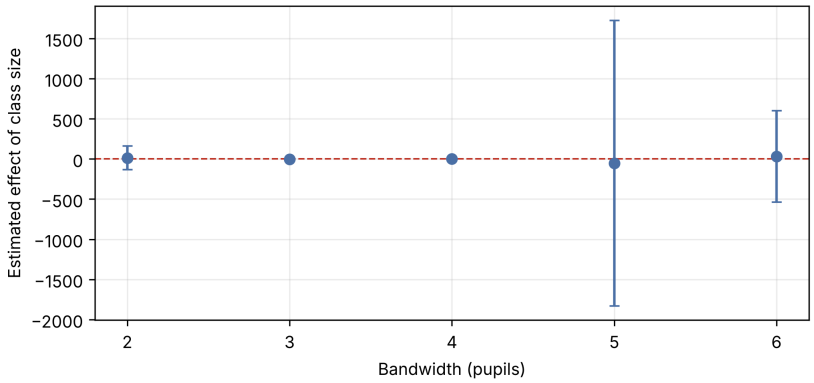


Fig. 4 Fixed-window discontinuity estimates and 95% confidence intervals by bandwidth, mathematics. The intervals at ±2, ±5 and ±6 extend far beyond the plotted range.

Table 9 Covariate balance at the discontinuities

	Bandwidth ±3			Bandwidth ±4		
Covariate	δ_1	SE	p	δ_1	SE	p
Pupil is male	−0.036	0.026	.177	−0.046	0.021	.029
Pupil age	−0.103	0.108	.336	−0.232	0.100	.020
Home educ. resources	0.072	0.286	.801	0.300	0.227	.186
missing indicator	0.004	0.029	.886	−0.000	0.021	.991
Parental education	−0.131	0.149	.379	−0.282	0.127	.026
missing indicator	0.009	0.025	.711	−0.002	0.020	.915
Teacher is male	0.205	0.151	.175	0.273	0.132	.038
Teacher holds master's	−0.126	0.125	.314	−0.093	0.125	.456
Teacher experience	6.489	3.680	.078	4.770	2.790	.087
% disadvantaged	−0.138	0.358	.700	−0.219	0.298	.463
School is urban	0.123	0.176	.486	0.183	0.145	.206
Unbalanced at 5%	0 of 11			4 of 11		

Estimates of δ_1 in Equation (5), with enrollment centred on each cut-off and data pooled across cut-offs. Standard errors clustered at the school level.

6.7 Class-level estimates

Table 10 reports the class-level least squares, first-stage and reduced-form estimates for the 218 single-class schools. Three features stand out. Least squares coefficients are negative in all five subjects and attenuate when PD is added, as in the pupil-level analysis. The first stage is strong with PD alone (0.780, $F = 84.4$) and remains conventionally strong when enrollment is controlled linearly (0.488, $F = 16.3$), comparable to the coefficients of 0.704 and 0.542 reported for Israeli fifth graders (Angrist & Lavy, 1999). And the reduced form is negative in every subject and becomes significant in all five once enrollment is controlled: a one-pupil increase in predicted class size is associated with a fall of 2.1 to 2.9 points in the class mean score.

Table 10 Class-level least squares, first stage and reduced form

	Least squares: class size			Predicted class size f	
Controls	(1) None	(2) PD	(3) PD, e	(4) PD	(5) PD, e
Class size (first stage)				0.780*	0.488*
				(0.085)	(0.121)
First-stage F				84.4	16.3
Mathematics	−1.957*	−1.372*	−1.762*	−1.418	−2.113*
	(0.554)	(0.482)	(0.545)	(0.811)	(0.974)
Physics	−1.732*	−0.994	−1.412*	−1.387	−2.385*
	(0.677)	(0.621)	(0.701)	(0.948)	(1.097)
Biology	−1.900*	−1.066	−1.539	−1.655	−2.865*
	(0.807)	(0.758)	(0.844)	(1.134)	(1.308)
Chemistry	−1.580*	−0.944	−1.277	−1.703	−2.681*
	(0.671)	(0.649)	(0.739)	(0.889)	(1.042)
Earth science	−1.961*	−1.235*	−1.767*	−1.396	−2.513*
	(0.671)	(0.619)	(0.688)	(0.917)	(1.043)

$N = 218$ classes (7,393 pupils) in single-class schools. Unweighted regressions of the class mean score; heteroscedasticity-robust standard errors in parentheses, combined across the five plausible values. Columns (1)–(3): coefficient on teacher-reported class size. Columns (4)–(5): coefficient on rule-predicted class size f, for class size (first stage) and for achievement (reduced form). PD: percentage of economically disadvantaged pupils (four categories); e: Grade 8 enrollment. In the ±5 discontinuity sample (12 classes) the first-stage coefficient is 0.530 (0.432; $F = 1.5$) with PD and 0.780 (0.245; $F = 10.1$) with PD and enrollment. *$p < .05$.

Table 11 reports the two-stage estimates. With PD alone (column 1) they lie between −1.8 and −2.2 points per pupil and only chemistry is significant. Under the linear enrollment control (column 2), the estimates are negative and significant in all five subjects, from −4.33 (SE = 2.04) in mathematics to −5.87 (2.58) in biology. Adding the quadratic term (column 3) increases the point estimates but doubles their standard errors, and none remains significant; the piecewise linear trend (column 4) produces estimates of −8 to −11 points, significant in biology, chemistry and earth science. In the ±5 discontinuity sample (columns 5 and 6) the estimates are close to zero, with standard errors larger than the estimates.

Two features of Table 11 qualify its significant entries. The magnitudes are large: the column (2) estimate in mathematics implies that a reduction of eight pupils would raise achievement by about 0.46 SD, more than twice the experimental benchmark (Krueger, 1999), and in class-mean units it is three to four and a half times the Israeli estimates. And the first stage weakens as soon as the enrollment control becomes more flexible: F falls to 4.4 under the quadratic control and to 5.0 under the trend, so that conventional inference in columns (3) and (4) is unreliable.

Table 12 therefore reports Anderson–Rubin sets for each class-level specification. Under the linear control the sets are bounded and exclude zero in all five subjects — in mathematics [−9.80, −0.46] — so the weak-identification-robust test confirms the sign of column (2) while leaving the magnitude anywhere between 0.006 and 0.13 SD per pupil. Under the trend the sets also exclude zero in every subject, but extend to −55 points per pupil or beyond, which carries no information about magnitude; under

Table 11 Class-level two-stage least squares estimates

	Full sample				±5 discontinuity sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Mathematics	−1.817	−4.333*	−6.645	−8.093	−0.814	−0.074
	(1.006)	(2.037)	(4.023)	(4.198)	(2.669)	(1.647)
Physics	−1.777	−4.890*	−7.779	−9.196	−1.009	1.050
	(1.168)	(2.202)	(4.562)	(4.821)	(5.499)	(3.203)
Biology	−2.121	−5.874*	−9.577	−11.212*	−0.921	1.704
	(1.396)	(2.577)	(5.319)	(5.621)	(6.784)	(4.082)
Chemistry	−2.182*	−5.496*	−8.979	−10.191*	−1.821	0.552
	(1.093)	(2.174)	(4.734)	(5.002)	(5.592)	(3.117)
Earth science	−1.789	−5.152*	−8.287	−9.896*	−2.299	0.340
	(1.127)	(2.128)	(4.525)	(4.828)	(6.125)	(3.391)
Controls	PD	PD, e	PD, e, $e^2/100$	Trend	PD	PD, e
First-stage F	84.4	16.3	4.4	5.0	1.5	10.1
Classes	218	218	218	218	12	12

Class size instrumented by f. Unweighted; heteroscedasticity-robust standard errors in parentheses, combined across the five plausible values. Trend: piecewise linear trend of Equation (7), with no other covariate. The ±5 discontinuity sample retains schools with enrollment in $[36, 45]$, $[76, 85]$ or $[116, 125]$. For comparison, the corresponding column (2) estimates for Israeli fifth graders are −0.275 in reading and −0.230 in mathematics, or −0.036 and −0.024 standard deviations of the class-mean distribution (Angrist & Lavy, 1999, Table IV); the Moroccan mathematics estimate is −0.109 standard deviations of that distribution (SD 39.8) and −0.058 of the weighted pupil-level distribution (SD 74.6). $^*p < .05$.

the quadratic control they exclude zero in three of five subjects. Under the bridge specification, which adds the cubic polynomial of the pupil-level design to the class-level model, the first-stage F falls to 2.9 and the sets are unbounded in every subject: with the cubic control the class-level design behaves as the pupil-level design does. Under the linear control, by contrast, the two designs differ — the pupil-level set, $[-5.75, 0.75]$, includes zero — a difference attributable to the pupil, teacher and school covariates, the weighting and the classes of multi-class schools that the pupil-level design adds.

The local component of the class-level design, which would make the estimates independent of parametric enrollment controls, cannot be implemented with these data (Table 13). The ±5 discontinuity sample contains 12 Moroccan classes, against 471 Israeli fifth-grade classes in the same windows (Angrist & Lavy, 1999), and only five of them lie above a cut-off; no sampled school enrolls between 41 and 45 pupils, so the first segment — the one at which the rule prescribes the largest reduction — contributes no comparison at all. The scarcity is not a consequence of the sample restrictions: of the 250 sampled schools, whatever the availability of class size, only 14 have an enrollment in the three intervals. The dummy-instrument estimates in the ±5 and ±3 samples rest on 8 to 12 observations for four to six parameters, and none of them is interpretable, including the two that are nominally significant.

Two checks complete the class-level analysis. Excluding the 12 schools whose recorded enrollment is inconsistent with the sampling frame (Section 4.3) leaves the

Table 12 Weak-identification-robust inference for the class-level specifications

	(1) PD	(2) Linear	(3) Quadratic	(4) Trend	(B) Cubic
First-stage π_1	0.780	0.488	0.301	0.333	0.273
	(0.085)	(0.121)	(0.143)	(0.149)	(0.160)
First-stage F	84.4	16.3	4.4	5.0	2.9
Anderson–Rubin 95% confidence sets					
Mathematics	$[-3.83,\ 0.23]$	$[-9.80,\ -0.46]$	$[-75.38,\ 0.28]$	$[-55.15,\ -1.75]$	unbounded
Physics	$[-4.07,\ 0.65]$	$[-10.58,\ -0.55]$	$[-84.78,\ 0.10]$	$[-64.27,\ -2.10]$	unbounded
Biology	$[-4.86,\ 0.78]$	$[-12.47,\ -0.72]$	$[-101.33,\ -0.50]$	$[-77.16,\ -3.09]$	unbounded
Chemistry	$[-4.36,\ 0.05]$	$[-11.46,\ -1.45]$	$[-103.43,\ -2.00]$	$[-73.34,\ -3.48]$	unbounded
Earth science	$[-4.00,\ 0.55]$	$[-10.84,\ -1.08]$	$[-91.83,\ -1.07]$	$[-67.80,\ -3.07]$	unbounded

$N = 218$ classes. Specifications (1)–(4) as in Table 11: PD; PD and linear enrollment; PD, enrollment and enrollment squared divided by 100; piecewise linear trend. (B) is a bridge specification with PD and a cubic polynomial in enrollment. Sets obtained by exact inversion of the Anderson–Rubin test with heteroscedasticity-robust variances combined across the five plausible values. Under (B) the set is the whole real line in mathematics and physics, and the union of two rays in biology ($(-\infty, 9.64] \cup [15.49, \infty)$), chemistry ($(-\infty, 0.51] \cup [31.51, \infty)$) and earth science ($(-\infty, 7.88] \cup [13.70, \infty)$). The 2SLS estimates under (B) range from -5.41 (4.21) in mathematics to -7.77 (5.04) in chemistry, none significant.

column (2) estimates of Table 11 essentially unchanged: -4.78 (2.20) in mathematics and between -5.50 and -6.60 in the science subjects, all significant. And the descriptive statistics of Appendix Table A1 show that the class-level sample resembles the full sample, and that the 12 classes of the discontinuity sample come from much smaller and more disadvantaged schools (mean enrollment 86 against 267, mean PD category 3.8 against 3.3), so that even a precise local estimate would describe a small and atypical segment of Moroccan lower secondary schools.

6.8 Robustness to the treatment of missing data

Table 14 reports a complete-case analysis, restricted to the 7,532 pupils with no imputed value on the Home Educational Resources scale or on parental education. The least squares coefficient in mathematics is -1.727 (SE = 0.466), against -1.759 (0.474) under imputation; the first-stage F is 3.84 against 3.94; and the two-stage estimate is -2.387 (3.902) against -2.858 (3.870). No conclusion is sensitive to the treatment of missing covariates, which is unsurprising given that neither the treatment nor the instrument is imputed.

6.9 Summary of the identification evidence

In the pupil-level design, ten two-stage estimates were obtained with an enrollment control, five under the linear and five under the cubic polynomial. None is significant; the first stage is strong under the linear control ($F = 26.0$) and weak under the cubic ($F = 3.9$), where the Anderson–Rubin set for mathematics is unbounded. In the discontinuity samples of that design, ten estimates at the reference bandwidths and four local polynomial estimates were obtained; none is significant and none has a first stage of conventional strength. The only significant estimate of the pupil-level design is the mathematics coefficient in column (4) of Table 4, obtained without an enrollment control, which the design does not permit.

Table 13 Discontinuity samples and dummy-instrument estimates

	±5 sample, PD	±3 sample, PD	±3 sample, no PD
Classes below / above the cut-off			
Segment 1 (cut-off 40)	3 / 0	1 / 0	1 / 0
Segment 2 (cut-off 80)	2 / 2	1 / 2	1 / 2
Segment 3 (cut-off 120)	2 / 3	1 / 3	1 / 3
Total, Morocco (Israel, fifth grade)	12 (471)	8 (302)	8 (302)
Two-stage least squares with dummy instruments			
Mathematics	−0.019 (1.039)	−1.176* (0.560)	−4.784* (2.249)
Physics	0.591 (1.940)	0.491 (1.638)	−4.424 (3.038)
Biology	0.531 (2.100)	1.518 (2.116)	−4.361 (3.594)
Chemistry	−0.113 (1.688)	0.481 (1.518)	−3.444 (2.429)
Earth science	−0.195 (1.787)	−0.155 (1.611)	−5.178 (3.093)
First-stage F (conventional)	1.5	1.5	3.2
Residual degrees of freedom	6	2	3

The ±5 (±3) sample retains schools with enrollment in $[36, 45]$, $[76, 85]$ or $[116, 125]$ ($[38, 43]$, $[78, 83]$ or $[118, 123]$). Instruments: z_{2s} and z_{3s}, indicating the upper half of the second and third segments; z_{1s} cannot be used because no school has an enrollment between 41 and 45. Controls: dummies for the first two segments, with or without PD. Heteroscedasticity-robust standard errors combined across plausible values; the first-stage F is reported in its conventional form because the robust covariance matrix of the two instruments is rank-deficient with so few observations. Israeli sample sizes from Angrist and Lavy (1999, Table VI). None of these estimates is interpretable; they are reported for completeness. $^{*}p < .05$.

Table 14 Sensitivity to the treatment of missing covariates, mathematics

	Median imputation (main)	Complete cases
N pupils	8,032	7,532
N schools	229	229
OLS, cubic enrollment control	−1.759 (0.474)	−1.727 (0.466)
First-stage π_1	0.286	0.283
First-stage F	3.94	3.84
2SLS, cubic enrollment control	−2.858 (3.870)	−2.387 (3.902)

Standard errors in parentheses, clustered at the school level, combined across the five plausible values. Complete cases exclude pupils with an imputed value on the Home Educational Resources scale or on parental education; the missingness indicators are dropped in that column because they are constant.

In the class-level design, the evidence leans towards a negative effect but depends on the enrollment control. Under the linear control, all five estimates are significant, the first-stage F is 16.3 and the Anderson–Rubin sets are bounded and exclude zero. Under the quadratic control and the piecewise linear trend, the first stage is weak and the sets, while excluding zero in eight of ten cases, are too wide to bound the effect. Under the cubic control the sets are unbounded. The local component of the design, which would make the result independent of the enrollment control, rests on

12 classes. No estimate in this study is therefore both significant and independent of a functional-form assumption about the direct effect of school size on achievement.

7 Discussion

7.1 What the study establishes and what it does not

It is worth separating the two explicitly, because the distinction is easily lost in summary.

The study establishes six things. Class size is negatively associated with achievement in all five subjects in Morocco, and the association survives adjustment for pupil, teacher and school characteristics in mathematics, at -1.76 points or -0.024 SD per pupil. Roughly a third of the raw association in mathematics, and about half in the science subjects, is accounted for by composition. The maximum-class-size rule predicts actual class size strongly in the unconditional sense, with a first-stage F of 87.4, but that relationship is largely a reflection of school size: conditioning on a cubic in enrollment reduces the first-stage coefficient from 0.846 to 0.286 and the F statistic to 3.9. In the class-level analysis, predicted class size is negatively and significantly related to achievement once enrollment is controlled linearly, and the weak-identification-robust sets exclude zero under the linear and piecewise linear controls. No discontinuity estimate is significant at any bandwidth examined, the first stage is weak throughout the discontinuity samples, and the narrow windows around the first three cut-offs are almost empty in Morocco. And the enrollment density is continuous at the cut-offs, so the Chilean manipulation problem does not arise here.

The study does not establish that reducing class size in Morocco would have no effect. Nor does it establish that the true causal effect is negative independently of how school size is modelled: the class-level sets exclude zero under the linear and piecewise linear controls, but under the cubic control they are unbounded at both the pupil and the class level, and under weak identification the two-stage estimator reproduces the sign of the least squares estimate mechanically. It does not establish the magnitude of the effect: the only bounded set that excludes zero, under the linear class-level specification, spans more than an order of magnitude, from 0.006 to 0.13 SD per pupil. It does not establish that the true effect is zero: the Anderson–Rubin set under the cubic specification is unbounded, which means the data are consistent with effects of any magnitude in either direction once that specification is adopted. It does not establish that class-size reduction would be an ineffective policy, which is a different proposition requiring cost as well as effect. And it does not establish that Maimonides-type designs are invalid in general, or that they would fail in every system resembling Morocco.

7.2 Why the rule yields little usable variation in Morocco

The weakness of the first stage under the required specification is not an accident of this sample. It follows from the interaction between a high maximum and a large-cohort enrollment distribution, and the magnitude involved can be computed from the institutional parameters alone.

The class size prescribed by a maximum of R is a saw-tooth function whose drop at the kth cut-off, from R pupils to $(kR+1)/(k+1)$, equals $(R-1)/(k+1)$ — approximately $R/(k+1)$ — and therefore falls rapidly in k. Table 15 reports the first few values for $R = 40$ together with the share of Moroccan schools in each range.

Table 15 Prescribed drop in class size at each discontinuity, $R = 40$

Transition	Enrollment at the cut-off	Prescribed drop (pupils)	Share of schools
$1 \to 2$ classes	40	19.5	3.5%
$2 \to 3$ classes	80	13.0	7.0%
$3 \to 4$ classes	120	9.8	7.0%
$5 \to 6$ classes	200	6.5	8.7%
$10 \to 11$ classes	400	3.5	4.8%

The last column is the share of the 229 sampled schools whose prescribed number of classes equals the lower value of the transition.

The median Moroccan middle school in the sample enrolls 245 Grade 8 pupils, and 72.5% of schools are prescribed five classes or more. These schools sit in the flat portion of the saw-tooth, where crossing a multiple of 40 implies a prescribed reduction of fewer than seven pupils — smaller, in the upper range, than the residual dispersion of reported class size around the predicted value, which has a standard deviation of 6.7. The signal the design relies on is therefore smaller than the noise in the variable it is meant to instrument, before any question of compliance arises. Section 4.4 noted that the observed dispersion combines genuine non-compliance with teacher reporting error and that the two cannot be separated here; for the present argument the distinction is immaterial, since both contribute to the same attenuation.

The contrast with the settings in which the design has produced significant estimates is arithmetic. In Israel, grade enrollment averaged 77.7 pupils with a median of 72 against a maximum of 40, so the typical school was prescribed two classes and sat at the steepest discontinuity, where crossing the cut-off implies a drop of twenty pupils; the rule accordingly explained about half the variation in class size (Angrist & Lavy, 1999). In the four European systems where the design has been applied to TIMSS, the maxima were lower still — 30 pupils in Hungary, Lithuania and Romania and 28 in Slovenia — and Grade 8 enrollments were concentrated well below 300 pupils, which again places most schools near the early discontinuities; the discontinuity analyses could accordingly draw on three to seven populated segments (Shen & Konstantopoulos, 2021). The Moroccan configuration reverses both terms. The same logic was already visible within the European results, where the instrument was weakest in Hungary, the country with the poorest compliance.

The windows around the first three cut-offs make the contrast concrete. Schools within five pupils of these cut-offs accounted for 471 of 2,019 Israeli fifth-grade classes, slightly fewer than a quarter (Angrist & Lavy, 1999). In Morocco they account for 12 of the 218 classes of the class-level sample, or 5.5%, with no class above the first cut-off (Table 13). Rule-based identification is strongest where the steepest discontinuities

are populated. In Morocco they are almost empty, so that the class-level estimates must rely on global controls for enrollment, whose functional form then decides the result (Section 6.7).

This suggests a limitation on the portability of rule-based class-size instruments that is testable before any estimation is attempted. The relevant quantity is the distribution of schools across the segments of the saw-tooth, which requires only the enrollment distribution and the statutory maximum, together with a measure of dispersion between predicted and actual class size. A design that concentrates its identifying variation in cut-offs where the prescribed drop is smaller than that dispersion cannot be expected to deliver a strong first stage under the enrollment controls the exclusion restriction requires. Whether this configuration recurs in other systems combining large cohorts with high maxima — a combination found in parts of the Middle East, North Africa and South Asia — cannot be established from a single country, and we make no such claim. We note only that the diagnostic is cheap, and that the present case provides evidence consistent with the concern.

7.3 Comparison with recent evidence

Placed against the recent literature, the Moroccan result is unexceptional. The clearest signal comes from the setting that founded the design: re-examining Israel with later cohorts, Angrist et al. (2019) found little evidence of effects comparable to those of the 1999 study and documented behavioural responses around the enrollment thresholds. Other rule-based studies are dispersed: effects indistinguishable from zero in Norway (Leuven et al., 2008), small effects on grade promotion in France (Gary-Bobo & Mahjoub, 2013), and positive effects on sixth-grade language scores in Japan (Akabayashi & Nakamura, 2014) — the latter obtained, as here, under a maximum of 40, which shows that a high maximum does not by itself preclude identification where compliance is close. The Japanese case is instructive precisely because it isolates the role of compliance from the role of the maximum.

The class-level estimates can be compared directly with the Israeli evidence (Angrist & Lavy, 1999). Under the linear enrollment control, the Moroccan first-stage coefficient of 0.49 is close to the Israeli 0.54, but the mathematics estimate corresponds to -0.11 standard deviations of the class-mean distribution per pupil, against -0.02 to -0.04 for Israeli fifth graders. A first stage of comparable strength thus yields an effect three to four and a half times larger. Two readings are possible: that class size matters more where classes are larger and resources scarcer (Altinok & Kingdon, 2012), or that a linear control does not absorb the direct effect of school size over an enrollment range several times wider than the Israeli one. Because identification is lost under the cubic control, the data cannot discriminate between them.

The international assessment literature points the same way. Woßmann and West (2006) found sizable effects in only two of eleven countries. Rule-based analyses of TIMSS and PIRLS have detected effects in a minority of systems (Konstantopoulos & Shen, 2016; Li & Konstantopoulos, 2016; Shen & Konstantopoulos, 2017), and in the most closely related study significant achievement effects appeared mainly in Romania in 2003, at 0.05 to 0.09 SD per pupil, while most other estimates were indistinguishable from zero and first-stage F statistics fell below 10 in most discontinuity samples (Shen

& Konstantopoulos, 2021). Read against that record, the present results extend a pattern rather than contradicting one.

The syntheses sharpen the point. Early meta-analytic work reported sizable benefits (Glass & Smith, 1979), but design-sensitive reviews find at most a small effect in reading and no robust effect in mathematics (Filges et al., 2018), and a recent correction for publication bias and model uncertainty leaves effects negligible outside Project STAR (Opatrný et al., 2025). The pupil-level estimate — imprecise and centred on a negative value — is what the corrected literature predicts; it is also, by construction, consistent with almost any hypothesis, which is why we draw no substantive conclusion from it. The class-level estimates under the linear control are, conversely, larger than anything the corrected literature would lead one to expect, which is a further reason to read them as fragile rather than as a discovery.

Among developing and middle-income systems the evidence is mixed. A positive instrumented coefficient has been reported in Bangladesh (Asadullah, 2005) and negative effects of larger classes in rural Bolivia (Urquiola, 2006), while the most recent study finds little learning gain from the class-size reductions India has financed over the ranges actually observed (Datta & Kingdon, 2023). The Indian case is the closest analogue in institutional terms. It also bears on the conjecture that class-size effects are larger in developing countries and smaller where teacher quality is higher (Altinok & Kingdon, 2012): Morocco is the more resource-constrained system on the indicators available here — a mean Grade 8 mathematics score of 378 against a centre point of 500 (von Davier, Kennedy, et al., 2024), and an average class of 36.5 pupils against roughly 23 to 26 in the four European countries — yet its estimates resemble those of Hungary and Slovenia rather than Romania. The present design cannot discriminate between a substantive failure of the conjecture and a simple lack of power, and the unbounded Anderson–Rubin set makes that explicit.

Within Morocco, the finding of Liouaeddine et al. (2017) deserves particular attention, because it is simultaneously a complement and a threat. Their report of a direct negative effect of Grade 8 enrollment on mathematics and science achievement is exactly the channel the exclusion restriction must rule out. If grade enrollment affects achievement other than through class size, the cubic control absorbs that channel only to the extent that the relationship is smooth; and whatever the control absorbs, it also removes from the instrument. The weak first stage documented here and the direct enrollment effect documented there are two readings of the same feature of Moroccan middle schools, and the two findings should be treated jointly. Finally, Lagdiri and Ouazzani Touhami (2026) report an optimal class size of 23 to 24 pupils using PISA 2022, which would imply that Moroccan classes averaging 36.5 pupils sit well above the optimum. That estimate identifies a different quantity — an optimum fitted across the observed distribution, rather than a local effect at administrative thresholds — so the two results are not strictly in conflict, but the tension indicates that the question of nonlinearity in the range above 35 pupils remains open.

7.4 Policy implications

The policy content of this study is modest by construction, and overstating it would contradict its own argument.

First, the study does not provide a sufficiently precise causal estimate to justify or to rule out a general class-size reduction in Moroccan lower secondary education. The adjusted association is negative and of a policy-relevant magnitude, and the class-level estimates point in the same direction. But an association is not a return, the class-level sign rests on a functional-form assumption about school size, and the confidence set under the cubic specification contains both zero and effects large enough to justify substantial expenditure. A ministry allocating a fixed budget cannot price the intervention from these data.

Second, the study documents substantial deviation from the existing maximum. A quarter of pupils are enrolled in classes above 40, the median class holds 38 pupils, and reported class size exceeds the rule-predicted value systematically among larger schools. Whatever the return to reducing the maximum, the return to enforcing the one already in force is a prior question, and it concerns a quarter of the Grade 8 population.

Third, the two findings are connected in a way that constrains evaluation. Non-compliance is both a policy problem and the reason the instrument is weak: a rule that binds would generate the variation needed to evaluate it. Strengthening enforcement would, as a by-product, make the effect of class size estimable in future cycles.

Fourth, the input most likely to determine the net effect of any reduction is not measured here. Smaller classes require more rooms and more teachers, and rapid expansion can dilute teacher experience enough to offset the gain (Jepsen & Rivkin, 2009). In a system where teacher recruitment is under strain, an evaluation on achievement alone is incomplete.

Fifth, and most practically, the evaluation of class-size policy in Morocco requires administrative data. Enrollment recorded at class formation and register-based class size would remove the two measurement problems of Sections 4.3 and 4.4 at once, and would allow the design applied here to be tested under the conditions for which it was developed.

8 Limitations

The first and governing limitation is weak identification. Under the cubic specification the first-stage F is 3.9, the Anderson–Rubin confidence set is unbounded, and the just-identified structure means that neither LIML nor Fuller offers any remedy. The class-level specifications that do exclude zero rely on linear or piecewise linear controls for enrollment whose adequacy over the Moroccan enrollment range cannot be tested, and their local counterpart rests on 12 classes. The instrumented estimates are reported rather than replaced by specifications selected on significance, but they are not point estimates of a causal effect and should not be cited as such.

Second, grade enrollment is measured at the start of the testing month rather than at class formation, which introduces measurement error into both the instrument and the running variable, with disproportionate effect near the cut-offs (Section 4.3). The direction of the resulting bias cannot be signed, and the public-use data cannot reconstruct enrollment at class formation.

Third, class size is teacher-reported rather than register-based, which attenuates the first stage and the least squares coefficient and inflates the apparent dispersion from the rule; reporting error and genuine non-compliance cannot be separated with these variables (Section 4.4).

Fourth, the public-use files contain no indicator of school sector, so schools bound by the public norm cannot be distinguished from those that are not. This pools two populations with potentially different compliance behaviour and limits both the interpretation of the first stage and the external validity of the estimates to the Grade 8 population as a whole.

Fifth, the fixed-window discontinuity analysis assumes a constant effect across segments of enrollment, which may not hold, and the covariate imbalance at ± 4 shows that the local randomisation property does not survive widening the window. One of three placebo cut-offs also rejects, which weighs against a causal reading of the discontinuity estimates.

Sixth, the exclusion restriction is untestable and the Moroccan evidence gives positive reason to doubt it (Liouaeddine et al., 2017).

Seventh, the class-level analysis is subject to three data constraints: it is restricted to the 218 schools in which a single class was sampled; the percentage of disadvantaged pupils is available only in four categories; and the piecewise linear trend is applied over an enrollment range much wider than the one for which it was proposed.

Finally, the study uses secondary data at a macro level and offers no information on what happens inside classrooms of different sizes. It therefore cannot distinguish between the possibility that Moroccan classroom practice does not exploit the potential of smaller classes (Blatchford et al., 2011) and the possibility that the design simply lacks power. Resolving that would require classroom observation alongside the survey data.

9 Conclusion

This paper asked whether the Moroccan maximum-class-size rule can identify the effect of class size on Grade 8 achievement, using the TIMSS 2023 national sample, a pupil-level instrumental-variable and discontinuity design, and a class-level design.

The empirical finding is a negative adjusted association between class size and achievement in all five subjects, significant in mathematics at -1.76 points per pupil, or -0.024 standard deviations, after adjustment for pupil, teacher and school composition and for a cubic in grade enrollment.

The identification finding is that this association cannot be given a causal reading that does not depend on how school size is modelled. In the pupil-level design, the instrument is strong in the unconditional sense, with a first-stage F of 87.4, but the conditioning that the exclusion restriction requires removes most of its content: under the cubic enrollment control the first-stage coefficient falls from 0.846 to 0.286, the F statistic to 3.9, and the Anderson–Rubin confidence set becomes unbounded. Its discontinuity component yields no significant estimate at any bandwidth between ± 2 and ± 6 or under any kernel and polynomial combination, and its first stage is weak

throughout. In the class-level design, the estimates are negative and their Anderson–Rubin sets exclude zero under linear and piecewise linear enrollment controls; but the first stage weakens and the sets widen as the control becomes more flexible, they become uninformative under the cubic control, and the discontinuity sample that would make the result independent of the control contains 12 classes. The enrollment density is continuous at the cut-offs and covariates are balanced in the narrowest window, so the failure is one of relevance rather than of design validity — but relevance is not a lesser requirement.

The interpretation follows directly and is narrower than either a null or a positive result. These data do not show that class size has no effect in Morocco, and the class-level evidence leans toward a negative effect; but they do not show what the effect is. An unbounded confidence set is not a finding of no effect but a statement that the evidence is uninformative about magnitude, and a set that excludes zero only under a particular functional form for school size is not a finding of an effect. The negative point estimates of the pupil-level design carry little weight in either direction, because weak-instrument bias reproduces the sign of the least squares estimate mechanically.

The contribution lies in the institutional diagnosis. The Moroccan configuration — a maximum of 40 against a median cohort of 245, which places the typical school at the seventh discontinuity where the rule prescribes a reduction of about five pupils, smaller than the dispersion of reported class size around the prediction, and which leaves the windows around the steepest discontinuities almost empty — illustrates a limitation of rule-based identification that is computable in advance from the enrollment distribution and the statutory maximum. We do not claim that the design fails wherever these features coincide; a single country cannot support that claim. We do suggest that the diagnostic is worth computing before the design is adopted, and that the many systems combining large schools with high caps are the ones for which it matters most. Establishing the effect of class size in such systems will require administrative data with enrollment recorded at class formation, or a source of exogenous variation other than the cap.

Declarations

- **Funding.** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- **Competing interests.** The author declares no competing interests.
- **Ethics approval and consent to participate.** Not applicable. The study is a secondary analysis of publicly available, fully anonymised data collected by the International Association for the Evaluation of Educational Achievement (IEA). The original data collection was carried out under the IEA's ethical and national approval procedures, and no further ethics approval was required.
- **Consent for publication.** Not applicable.
- **Data availability.** The TIMSS 2023 international database analysed in this study is publicly available from the TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College, and the IEA at https://timss2023.org/data.

- **Materials availability.** Not applicable.
- **Code availability.** The analysis code used in this study is available from the author upon reasonable request.
- **Author contribution.** The author is solely responsible for the conception and design of the study, data analysis and interpretation, and drafting and revision of the manuscript, and approved the final version.

Appendix A Class-level descriptive statistics

Table A1 Descriptive statistics for the class-level sample

			Quantiles				
Variable	Mean	SD	0.10	0.25	0.50	0.75	0.90
A. Class-level sample (218 classes in 218 schools)							
Class size	35.9	6.3	26.0	33.0	37.0	40.0	43.0
Grade 8 enrollment	267.0	156.3	82.4	153.2	247.0	344.2	485.5
Disadvantaged pupils (1–4)	3.3	1.0	1.0	3.0	4.0	4.0	4.0
Pupils tested	33.9	6.4	25.0	31.0	34.0	38.0	41.0
Mean mathematics score	373.9	39.8	329.7	346.4	369.7	394.5	423.4
Mean science score	317.8	49.7	261.7	281.3	307.0	346.5	391.0
B. ±5 discontinuity sample (12 classes; enrollment 36–45, 76–85, 116–125)							
Class size	33.3	7.7					
Grade 8 enrollment	86.0	34.7					
Disadvantaged pupils (1–4)	3.8	0.9					
Pupils tested	32.7	8.0					
Mean mathematics score	376.8	41.8					
Mean science score	318.6	56.4					

Unweighted class-level statistics. Mean scores are class means averaged over the five plausible values; the science score is the overall TIMSS science scale. Disadvantaged pupils: TIMSS category of the percentage of economically disadvantaged pupils (1: 0–10%; 2: 11–25%; 3: 26–50%; 4: more than 50%). Quantiles are reported for panel A only. For comparison, Israeli fifth-grade classes had a mean size of 29.9 (SD 6.5) and a mean grade enrollment of 77.7 (SD 38.8), with a 90th percentile of 128 (Angrist & Lavy, 1999, Table I).

References

Akabayashi, H., & Nakamura, R. (2014). Can small class policy close the gap? an empirical analysis of class size effects in Japan. *The Japanese Economic Review*, *65*(3), 253–281, https://doi.org/10.1111/jere.12017

Altinok, N., & Kingdon, G. (2012). New evidence on class size effects: A pupil fixed effects approach. *Oxford Bulletin of Economics and Statistics*, *74*(2), 203–234, https://doi.org/10.1111/j.1468-0084.2011.00648.x

Angrist, J.D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575, https://doi.org/10.1162/003355399556061

Angrist, J.D., Lavy, V., Leder-Luis, J., Shany, A. (2019). Maimonides' rule redux. *American Economic Review: Insights*, *1*(3), 309–324, https://doi.org/10.1257/aeri.20180120

Asadullah, M.N. (2005). The effect of class size on student achievement: Evidence from Bangladesh. *Applied Economics Letters*, *12*(4), 217–221, https://doi.org/10.1080/1350485042000323608

Bijou, M., & Bennouna, N. (2018). Dépenses publiques éducatives et performance scolaire au Maroc : Une analyse multiniveaux à partir des données TIMSS 2015. *HAL Working Paper hal-01689120*, ,

Blatchford, P., Bassett, P., Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction. *Learning and Instruction*, *21*(6), 715–730, https://doi.org/10.1016/j.learninstruc.2011.04.001

Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: Evidence from TIMSS. *Large-scale Assessments in Education*, *4*, 12, https://doi.org/10.1186/s40536-016-0030-6

Chatri, A., Chahbi, O., Snihji, M. (2021). The multilevel analysis of students' achievement: Evidence from Morocco. *African Development Review*, *33*(1), 117–129, https://doi.org/10.1111/1467-8268.12497

Cour des comptes (2018). *Référé du premier président sur les conditions de la rentrée scolaire 2016/2017* (Tech. Rep.). Rabat: Cour des comptes du Royaume du Maroc. (In Rapport annuel de la Cour des comptes au titre de 2016 et 2017)

Datta, S., & Kingdon, G.G. (2023). Class size and learning: Has India spent too much on reducing class size? *The World Bank Economic Review*, *37*(1), 24–48, https://doi.org/10.1093/wber/lhac025

Ehrenberg, R.G., Brewer, D.J., Gamoran, A., Willms, J.D. (2001). Class size and student achievement. *Psychological Science in the Public Interest*, *2*(1), 1–30, https://doi.org/10.1111/1529-1006.003

Filges, T., Sonne-Schmidt, C.S., Nielsen, B.C.V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, *14*(1), 1–107, https://doi.org/10.4073/csr.2018.10

Fishbein, B., Siegel, P., Atasever, U., Cooney, D., Gonzalez, G. (2024). Standard error estimation in TIMSS. M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report (methods and procedures)* (chap. 13). TIMSS & PIRLS International Study Center, Boston College.

Fredriksson, P., Öckert, B., Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics*, *128*(1), 249–285, https://doi.org/10.1093/qje/qjs048

Gary-Bobo, R.J., & Mahjoub, M.-B. (2013). Estimation of class-size effects, using "Maimonides' rule" and other instruments: The case of French junior high schools. *Annals of Economics and Statistics*, *111/112*, 193–225, https://doi.org/10.2307/23646331

Glass, G.V., & Smith, M.L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1*(1), 2–16, https://doi.org/10.3102/01623737001001002

Hoxby, C.M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*(4), 1239–1285, https://doi.org/10.1162/003355300555060

Ibourk, A. (2013). Determinants of educational achievement in Morocco: A micro-econometric analysis applied to the TIMSS study. *International Education Studies*, *6*(12), 25–36, https://doi.org/10.5539/ies.v6n12p25

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959, https://doi.org/10.1093/restud/rdr043

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources*, *44*(1), 223–250, https://doi.org/10.3368/jhr.44.1.223

Konstantopoulos, S., & Shen, T. (2016). Class size effects on mathematics achievement in Cyprus: Evidence from TIMSS. *Educational Research and Evaluation*, *22*(1–2), 86–109, https://doi.org/10.1080/13803611.2016.1193030

Krueger, A.B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532, https://doi.org/10.1162/003355399556052

Krueger, A.B., & Whitmore, D.M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results. *The Economic Journal*, *111*(468), 1–28, https://doi.org/10.1111/1468-0297.00586

Lagdiri, Y., & Ouazzani Touhami, N. (2026). Optimal class size and learning outcomes in Morocco: A normative-causal approach using PISA 2022 data. *International Journal of Educational Development*, *121*, 103523, https://doi.org/10.1016/j.ijedudev.2026.103523

Lazear, E.P. (2001). Educational production. *The Quarterly Journal of Economics*, *116*(3), 777–803, https://doi.org/10.1162/00335530152466232

Lee, D.S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355, https://doi.org/10.1257/jel.48.2.281

Leuven, E., Oosterbeek, H., Rønning, M. (2008). Quasi-experimental estimates of the effect of class size on achievement in Norway. *The Scandinavian Journal of*

Economics, *110*(4), 663–693, https://doi.org/10.1111/j.1467-9442.2008.00556.x

Li, W., & Konstantopoulos, S. (2016). Class size effects on fourth-grade mathematics achievement: Evidence from TIMSS 2011. *Journal of Research on Educational Effectiveness*, *9*(4), 503–530, https://doi.org/10.1080/19345747.2015.1105893

Liouaeddine, M., Bijou, M., Naji, F. (2017). The main determinants of Moroccan students' outcomes. *American Journal of Educational Research*, *5*(4), 367–383, https://doi.org/10.12691/education-5-4-5

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714, https://doi.org/10.1016/j.jeconom.2007.05.005

Médias24 (2022). *Rentrée scolaire 2022-2023 : 30 élèves par classe en première et deuxième années du primaire.* (20 July)

Mourji, F., & Abbaia, A. (2013). Les déterminants du rendement scolaire en mathématiques chez les élèves de l'enseignement secondaire collégial au Maroc. *Revue d'économie du développement*, *21*(1), 127–158,

Opatrný, M., Havránek, T., Havránková, Z., Ščasný, M. (2025). Publication bias and model uncertainty in measuring the effect of class size on achievement. *Journal of Labor Economics*, , https://doi.org/10.1086/737989 (Advance online publication)

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3–15, https://doi.org/10.1177/096228029900800102

Shen, T., & Konstantopoulos, S. (2017). Class size effects on reading achievement in Europe: Evidence from PIRLS. *Studies in Educational Evaluation*, *53*, 98–114, https://doi.org/10.1016/j.stueduc.2017.04.001

Shen, T., & Konstantopoulos, S. (2021). Estimating causal effects of class size in secondary education: Evidence from TIMSS. *Research Papers in Education*, *36*(5), 507–541, https://doi.org/10.1080/02671522.2019.1697733

Siegel, P., & Foy, P. (2024). Sample design. M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 technical report (methods and procedures)* (chap. 3). TIMSS & PIRLS International Study Center, Boston College.

Stock, J.H., Wright, J.H., Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518–529, https://doi.org/10.1198/073500102288618658

TIMSS, & Center, P.I.S. (2023a). *TIMSS 2023 school questionnaire: Grade 8.* Boston College.

TIMSS, & Center, P.I.S. (2023b). *TIMSS 2023 teacher questionnaire: Mathematics, grade 8.* Boston College.

Todd, P.E., & Wolpin, K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33, https://doi.org/10.1111/1468-0297.00097

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *The Review of Economics and Statistics*, *88*(1), 171–177, https://doi.org/10.1162/003465306775565729

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, *99*(1), 179–215, https://doi.org/10.1257/aer.99.1.179

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *TIMSS 2023 technical report (methods and procedures).* TIMSS & PIRLS International Study Center, Boston College.

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

Woessmann, L. (2005). Educational production in Europe. *Economic Policy*, *20*(43), 446–504, https://doi.org/10.1111/j.1468-0327.2005.00144.x

Woßmann, L., & West, M.R. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, *50*(3), 695–736, https://doi.org/10.1016/j.euroecorev.2004.11.005