Research Group on Human Capital
Working Paper Series

Nonlinear Class Size Effects on Cognitive and Noncognitive Development of Young Children

Working Paper No. 18-01

Marie Connolly and Catherine Haeck

November 2021 (revised version)

Groupe de recherche sur le
CAPITAL HUMAIN
ESG UQÀM

https://grch.esg.uqam.ca/en/working-papers-series/

Nonlinear Class Size Effects on Cognitive and Noncognitive Development of Young Children

Marie Connolly and Catherine Haeck*
Research Group on Human Capital, Université du Québec à Montréal

November 2021

Abstract

We estimate the nonlinear impact of class size on student achievement by exploiting regulations that cap class size at 20 students per class in kindergarten. Based on student-level information from a previously unexploited and unique large-scale census survey of kindergarten students, this study provides clear evidence of the nonlinearity of class size effects on development measures. While the effects are largest on cognitive development, class size reductions also improve noncognitive skills for children living in disadvantaged areas. These findings suggest that sizeable class size reductions targeted at disadvantaged areas would achieve better results than a marginal reduction across the board.

* *Catherine Haeck:* haeck.catherine@uqam.ca. We gratefully acknowledge financial support from the Fonds de recherche du Québec - Société et Culture. We thank David Lee, Alex Mas, Fabian Lange, Daniel Parent, Dalibor Stevanovic, Philip Merrigan, and also participants at the CEA, SCSE and SOLE meetings, the Boise State University seminar, the Université de Sherbrooke seminar, the GRCH workshop at the Université du Québec à Montréal, the IRLE visitors' workshop at UC Berkeley, and the JOLE Special Volume authors conference for their comments. We also want to thank the Institut de la statistique du Québec research data centre (CADRISQ) for great support throughout this project. David Lapierre provided skillful research assistance. The analysis is based on the Québec Survey of Child Development in Kindergarten restricted-access Micro Data Files available at the CADRISQ. All computations on these microdata were prepared by the authors who assume responsibility for the use and interpretation of these data.

1 Introduction

This article evaluates the nonlinear effects of class size on student achievement by exploiting regulations that set maximum class size at 20 pupils per class in kindergarten, adapting Angrist and Lavy's (1999) and Fredriksson et al.'s (2016) causal identification strategies to estimate nonlinearities. Mean effects may be hiding important variations, which has important implications for education policy and cost-benefit analyses. Our main contribution is to provide the first estimates of nonlinear effects of class size on cognitive and noncognitive measures of child development for kindergarten students using unique student-level quasi-administrative data. We carefully show that class size effects vary along the class size distribution such that the marginal effect of class size reduction decreases with class size. Focusing on kindergarten students is relevant because early-life interventions have long-lasting effects (Heckman, 2006), and because causal identification is facilitated by the fact that class composition manipulation is much less likely at the kindergarten level since school teachers and administrators know very little of the child before school starts.

Class size reduction is a popular educational reform. It is also costly, making it important to have good estimates of its benefits to be able to gauge its effectiveness vis-à-vis other education policies (Krueger, 2003). The benefits of smaller classes on student achievement[1] have been documented in a number of studies using credible empirical strategies, but the debate remains ongoing (Hanushek, 1999, 2002). Identification is complicated by the fact that class size is not random in most settings: pupils with learning or behavioral disabilities may be placed in smaller classes, and children of high-income parents and of higher ability may also have access to schools with more resources and smaller classes. Theory also suggests that optimal class sizes are larger for better-behaved students. Models are often based on disruptions, with turbulent students in the classroom disrupting teaching and negatively affecting student learning (Lazear, 2001). The larger the class, the more likely disruptions will be. Lazear's model is inherently nonlinear: the probability of disruption increases at a

[1]Some studies have also been interested in class size effects on noncognitive measures of development, such as disruptiveness, inattentiveness, motivation, effort, or self-confidence (Dee and West, 2011; Ding and Lehrer, 2011; Chetty et al., 2011).

faster rate for small classes than it does for large classes, in which disturbances are already more common. In consequence, the effect of reducing class size depends on the size of the class, and class size reduction has a larger impact in smaller classes.

Yet the overwhelming majority of studies on class size in primary school presents estimates based on linear models, meaning that the estimated effect of reducing class size is constant over the class size distribution. Exceptions include Hoxby (2000), Urquiola (2006), and Hojo (2013), who allow for nonlinearities, but without specifically testing for them.[2] Urquiola (2006) looks at the effects of class size reductions on third graders' achievement in Bolivia, where the maximum class size is 30. His comparison of two empirical strategies leads him to conclude that his evidence is consistent with nonlinearities in the class size effects, though without estimating the nonlinearities directly. Hojo (2013) uses a piecewise-linear function in a setting where maximum class size is 40. Using data on mathematics and science test scores of around 4,500 fourth graders in Japan, he finds that class size reduction generates a positive impact on test scores when the reduction occurs below 23. The relevance of these estimates for educational systems with overall much smaller classes on average, such as the United States where the average class size at the elementary school level was 21 in 2017 (OECD, 2017), is less clear. Hoxby (2000) uses school- and district-level data from Connecticut to estimate the impact of class size. She uses the natural logarithm of class size to account for nonlinearity. She finds that class size reduction does not improve September test scores in fourth and sixth grades. Class size is measured in prior school years and assumes that children do not change schools. In her setting, class size rules vary between districts which, as she mentions, implies that parents with unobserved characteristics positively related to test scores may choose to buy a residence in a district with smaller class sizes.

To conduct our analysis, we use previously-unexploited student-level data on more than 80 percent of kindergarten pupils in the Canadian province of Québec in 2012, a context that is comparable to the American one. We first present a conceptual framework that relates

[2]There are also a few studies using postsecondary education data. Bandiera et al. (2010) estimate heterogeneous effects of class size on test scores, but for university students. Bedard and Kuhn (2008) also focus on higher education, this time looking at the nonlinear effects of class size on teacher evaluations.

class size to student development through an education production function. We then show evidence on the nonlinearity of class size effects, adapting Angrist and Lavy's (1999) and Fredriksson et al.'s (2016) empirical approaches. Our fuzzy RD approach is based on the fact that class size is capped at 20 in Québec, a Maimonides-type setting that induces jumps along the school enrollment distribution. Since jumps differ in magnitude, we are able to capture nonlinear patterns in the relationship between class size and development scores by comparing the marginal effects at different discontinuities.

Our setting is in line with Project STAR (Krueger, 1999; Krueger and Whitmore, 2001; Chetty et al., 2011) in the sense that we study class size variations between nine and 24 students in kindergarten, while Angrist and Lavy (1999) look at variations between 20 and 40 students in third to fifth grades. In contrast with Hoxby (2000), the Québec class-size rule is the same for everyone across the entire province, thereby eliminating a possible source of bias. Also, because we have student-level data on kindergarten children, we observe the actual class size of the child and relate it directly with their development several months after starting school. Finally, in light of recent findings by Angrist et al. (2017) and Fredriksson et al. (2016), we also make sure our estimates are not biased by teachers' strategic behavior or bunching around enrollment thresholds.

Our results bring novel insights to the impact of class size on child development. We also find that, on average, class size reduction increases both cognitive and noncognitive scores for kindergarten children, but the magnitude of the average effects remains modest and comparable to Krueger (1999). Our main contribution is to confirm the presence of nonlinearities in the effect of class size on development scores, which opens the door for resource reallocation with a net gain. Finally, we also find evidence that children living in precisely-defined disadvantaged geographic areas benefit more from class size reduction, the benefits on other students being marginal and often not different from zero. These findings suggest that it is possible to design a cost-neutral intervention through class size reallocation that would have the potential to reduce inequalities between children.

From a policy standpoint, our findings suggest that while class size reduction policies always imply increased expenditures, they may not always trigger cognitive and noncognitive

gains for students. Large and targeted class size reduction policies in disadvantaged areas are far more efficient than modest universal reduction policies. Such targeted policies could help level the playing field and promote more equal opportunities, as argued by Piketty and Valdenaire (2006) in their study of class size in France with a focus on disadvantaged areas. From a budgetary standpoint, our findings further suggest that large reductions in a limited number of classes could be financed by marginal increases in the vast majority of schools not experiencing high poverty rates: such a policy could therefore be cost-neutral.

The rest of the paper is organized as follows. We start with a conceptual framework of class size effects in Section 2. We follow with a presentation of the institutional context and of the data set in Sections 3 and 4. Section 5 presents the identification strategy while Section 6 discusses our findings. The final section concludes.

2 Class size effects: a conceptual framework

The empirical literature on class size effects has mostly tried to answer the question of whether class size has an effect on test scores or some other measure of child development or educational outcome. In a typical model, test scores are explained by class size and other covariates using a linear model. The identification of a causal effect has been a central preoccupation, given the potential endogeneity of class size. But the functional form of the relationship between educational outcome and class size is also important. In particular, when one thinks about resource allocation, a linear homogeneous relationship would imply that reallocating teachers across schools or districts would not matter if the number of teachers is constant since the average class size would also remain constant. In such a linear context, class size reductions can only lead to an average positive outcome if the average class size decreases or if there are heterogeneous effects. However, nonlinearities–and heterogeneous effects–give rise to more interesting optimization strategies. In this context, even with a fixed average class size, a large reduction in class size for a small number of students compensated by small class size increases for the majority would lead to an overall gain. In other words, an internal solution to the optimization problem of class size requires

some nonlinearities in benefits (or costs).

The question thus becomes whether we can detect nonlinearities in the test scores-class size relationship. In doing so, the functional form is key, as is the choice of the variable used to measure the educational input and its units.

We start by conceptualizing, in very general terms, the education production function that we have in mind. Suppose the education attained, call it e, is a function of some inputs: $e = f(inputs)$. The questions we asked above are thus what functional form the $f(.)$ function takes, but also what are the relevant inputs of the education production function. Krueger (2002, p. 7) wrote: "Aside from the opportunity cost of students' time, teachers are the most important, and most costly, factor of production in education." A natural input would then be teacher's instruction time, perhaps more accurately per-student instruction time, which would mean that an input of interest is $T = NumTeachers/NumStudents$. What we are interested in, for optimization purposes, is thus how $f(.)$ varies with T.

Note however that the class size models estimated in practice are of the form $e = f(ClassSize, other\ inputs)$, but that class size is the reciprocal of T ($ClassSize = NumStudents/NumTeachers = 1/T$), meaning that we estimate $e = f(1/T, other\ inputs)$. If we forget about the other inputs for the moment, our descriptive evidence (presented in Section 4) leads us to think that the function $f(1/T)$ is convex, and that we are in the decreasing portion of it, meaning that f', the first derivative, is negative, and f'', the second derivative, is positive. The class size literature also supports the idea that $f'(1/T) < 0$, since the average marginal effect of increasing class size documented is generally negative, but the sign of the second derivative remains an open question since nonlinearities have not been the focus of many studies.

If we go back to our educational production function, the question becomes how does educational output vary when the input T changes? The first derivative of f with respect to T is $f'(1/T)(-1/T^2)$, and it is positive. The second derivative is $f''(1/T)(-1/T^2)^2 + f'(1/T)(2/T^3)$. The sign of the second derivative is of particular interest since it informs us about nonlinearities: if it is positive, then the relationship between educational output and input T is convex, and reallocating teachers (or teaching units) from large classes to small

classes would lead to improved outcomes. We will use our empirical work to get at a back-of-the-envelope investigation of the sign of this second derivative, which is equivalent to the sign of this expression: $f''(1/T) + f'(1/T)(2T) = f''(ClassSize) + f'(ClassSize)(2/ClassSize)$. We view this contribution as a way to frame the empirical research on class size effects, especially in a resource-allocation context.

3 Institutional background

We present here some relevant institutional background on kindergarten and schools in Québec. The general context is quite similar to most of the United States. Children start kindergarten at age five: a child must be five years old by September 30 of the entering school year. Schools are typically divided between elementary schools, covering kindergarten through grade six, and secondary schools, for grades seven and up. Elementary schools are on average smaller, and thus located closer to home. Children attend school 182 days a year, Monday to Friday, roughly from September to June, inclusively. Schedules vary from school to school, but total instruction time is usually just under five hours each day in kindergarten, not including lunch. Before- and after-school care, as well as lunch supervision, is subsidized by the Québec government, with a rate currently (and in 2012) set at C$8 a day. Most children attend public schools, for which enrollment is free: 93% of kindergarten-level students in 2011-2012 were in public schools. Education is under provincial jurisdiction, not federal. Public spending on elementary and secondary education was 3.9% of GDP in 2009-2010, compared to 4% in the United States (MEESR, 2015). Until 2020, public schools in Québec were organized by school boards. Boards, or districts, were divided in parallel boards according to language: there were 60 French-language boards, nine English-language boards, and one board serving both French and English schools, located in a remote area. Two additional boards regrouped Indigenous communities from northern Québec. Most children attend school in French: 90% of students in 2011-2012 for kindergarten.[3] Children are assigned to a public school based on the location of their residence. If they live within a

[3]Most statistics in this section come from the Databank of Official Statistics on Québec, available at www.bdso.gouv.qc.ca.

school's catchment area, they are guaranteed a spot. Parents also have the option to enroll their child to another school within the same school board as availability permits. Changing school board is more complicated and rarely done.

Québec generously subsidizes early-childhood care, either directly funding public low-cost daycare centers, or through refundable tax credits available for parents sending their child to a private daycare. Those subsidies stop when the child reaches kindergarten age, at which point the vast majority of children switch to the public elementary school system. Pre-kindergarten programs, also called junior kindergarten or four-year-old kindergarten in other contexts, have been offered in targeted schools from high deprivation areas as early as 1978 (Government of Québec, 2012). These programs were offered part-time (morning or afternoon, five days per week) on a voluntary basis. Parents had the option to send their child to pre-kindergarten if they lived in the school's catchment area and their child was four years of age by September 30 of the school year. One of the goals of these programs was to facilitate the transition to school for low socioeconomic status children. In school year 2012-2013, our study period, $81,541$ students were enrolled in kindergarten, but only $16,910$ of these students had attended part-time pre-kindergarten programs in a school setting the previous year (roughly 21%). Since 2013, pre-kindergarten is now offered full-time in a larger number of schools, and the network is expanding.

Maximum class size is regulated by the Québec government in its agreement with the school teachers. The rule is therefore the same in all schools across the province. In Hoxby (2000), the rule varied by school district with a modal maximum class size of 25. While parents in Connecticut could strategically choose their area of residence to send their child to a school with smaller classes, parents in Québec do not have that option. In fact, they are generally completely unaware of the actual class size until the start of the school year. For kindergarten, the maximum class size has been set at 20 since year 2000. Private schools are not subject to this class size cap; the average class size in private schools is higher than in public schools. This cap at 20 creates a situation that is similar to that in most of the United States. Hoxby (2000) reports that most school districts in the United States have maximum class size rules ranging between 20 to 30 students. A majority of states have caps, or no

hard cap but tie funding to class size. For example, California provides additional funding for classes with no more than 20 students in kindergarten (Gilraine et al., 2018). New York sets maximum class size at 20 as well, Texas has a cap of 22, and Florida uses 18.[4]

4 Data

Our data come from the 2012 Québec Survey of Child Development in Kindergarten (QSCDK), a large-scale census survey conducted by the Institut de la statistique du Québec (ISQ), the provincial equivalent to Statistics Canada (Institut de la statistique du Québec, 2013a). The QSCDK 2012 collected information on more than 80% of students enrolled in kindergarten in academic year 2011-2012 (Simard et al., 2013). The goal of this survey was to provide a detailed picture of the development of Québec children. The QSCDK has not been previously exploited to look at the effects of class size. In fact, only a couple of studies in the field of public health have used the QSCDK data (Laurin et al., 2015, 2018), making this paper the first study in economics to avail itself of these rich data. Overall, 98% of Québec's kindergarten students are targeted by the survey. Identified at-risk students (students with learning difficulties or special needs) registered in classes with 50% or more at-risk students are excluded from the QSCDK target population, as well as schools catering specifically to special needs students. Also excluded from the target population are students in schools under federal jurisdiction (mostly schools on First Nations reserves) or in schools in Québec's two Indigenous school boards, located in the northern areas of the province. Participation rate is extremely high: 69 out of 70 school boards and 88% of Québec's schools participated in the survey, for a combined response rate at the student level of more than 80%.

Five domains of development are measured in the QSCDK: (1) cognitive and language development, (2) social competence, (3) emotional maturity, (4) communication skills and general knowledge, and (5) physical health. These five domains are assessed using the Early Development Instrument (EDI) developed at McMaster University in Ontario, Canada (Janus and Offord, 2007). The questionnaire filled out by teachers, online in 95% of the cases,

[4]To see a map of the United States along with class size information, look at Education Week's infographics here: https://www.edweek.org/ew/section/infographics/13class_size_map.html.

is designed to assess the strengths and difficulties of children regarding the five domains.

In the EDI, the child's development is measured using a total of 104 questions (Yes/No/Don't know) split into three sections. Additional questions are included to gather information on child characteristics (15 questions), disabilities (17 questions), early childhood experiences (13 questions), and specific abilities (seven questions). The cognitive and language development domain refers to the child's competence in mathematics, reading and writing. It is measured, among other things, by the child's ability to recognize and use numbers from 1 to 10, read and write simple words or sentences, compare numbers, and identify concepts of time. Communication skills are measured by the ability of the child to listen, tell a story, pronounce clearly, and communicate their needs clearly. Social competence refers to the ability to play with others, follow the rules, act responsibly, adapt to change, and be autonomous. Emotional maturity captures whether the child has a tendency to hurt others physically or mentally, help others when in need, wait, and be patient, calm, and attentive. Additional information about the EDI can be found in Janus and Offord (2007). The EDI has been used not only in Canada, but also in Australia, the United States, and England (e.g., Janus et al., 2011; Brinkman et al., 2013), and has been shown to accurately measure the state of development of children. Janus et al. (2011) specifically test the validity of the measure, and show results against the well-known PPVT in four countries. Furthermore, recent research on the EDI shows that it strongly predicts children's literacy and numeracy assessments at ages eight, 10, and 12 (Brinkman et al., 2013). The questionnaire is provided in the online Appendix. While the EDI is not a high-stakes exam, it is a detailed measure with proven external validity. The advantage of not being a test, and definitely not a high-stakes test, alleviates concerns that can come from high-stakes testing (Lazear, 2006; Rockoff and Turner, 2010) and from score manipulation (Angrist et al., 2017).

One may be concerned by the fact that the EDI relies solely on teachers' answers. Forget-Dubois et al. (2007) specifically test the validity of the EDI against more traditional test-based measures administered by trained professionals in a longitudinal survey of Québec children. They find that the EDI, especially its cognitive component, predicts later school achievement in grade one over and above the cognitive assessments and direct school readi-

ness tests included in their survey, namely the PPVT-III (Dunn and Dunn, 1997), the Block Design subtest of the Wechsler Preschool and Primary Scale of Intelligence–Revised (Wechsler, 1989), the Number Knowledge Test (Okamoto and Case, 1996) and the Visually Cued Recall task (Zelazo et al., 2002). Their assessment concludes that the cognitive and language development score of the EDI is one of the best predictors of school achievement among the measures available in their study. They further conclude that teachers' assessment using the EDI questionnaire in kindergarten can be almost as effective as a formal battery of tests that rely upon trained professionals. However, the EDI has some limits which must be acknowledged. While all teachers in Québec have completed a four-year-long university degree, combined with several months of training on school premises and are considered highly qualified, assessments may vary across teachers. To ensure the EDI is completed carefully by teachers, the ISQ ensures that a substitute teacher is hired to allow each teacher to answer the questionnaires in a separate room during the teacher's working hours. Teachers take on average 15 to 20 minutes to answer each student's questionnaire and time is monitored by the Québec Statistical Institute who runs the survey. Finally, teachers answer the questionnaires after having observed the child for several months.

We use as outcomes the cognitive, social, emotional, and communication measures computed from the various Yes/No questions and provided in the data set. Physical health measures are used as controls. We also compute three additional development scores that we sometimes use as outcomes. First, we extract from the cognitive development measure the subquestions related to math abilities and those to reading and writing abilities.[5] Second, we create a single index of development based on the four core measures of the EDI equally weighted.[6] This allows us to capture the overall development of the child, but places a lot of weight on noncognitive development.

In addition to the developmental measures, the QSCDK provides information at the school and student levels, which we also mostly use as control variables. At the student

[5]As can be observed in the survey questionnaire provided in the online Appendix, subquestions 8 to 23 in Section B correspond to reading and writing abilities, while subquestions 25 to 33 correspond to math abilities. We keep only students for which all of the answers related to the subquestions are not missing.

[6]Equal weights were selected based on the results of a principal component analysis.

level we have information on gender, age in months, mother tongue, place of birth, number of years since child immigrated to Canada, quintile of social and material deprivation[7] of residential area, whether the child attended childcare or attended pre-kindergarten in a school setting (also called four-year-old kindergarten or junior kindergarten in other contexts), has social or learning disabilities (physical limitations, visual deficiencies, hearing deficiencies, problems at home, chronic health conditions, dental issues), and has required help from a non-teaching professional at school (nurse, speech therapist, psychoeducator, social worker, psychologist). School-level variables include total kindergarten enrollment, class size for each group, administrative region, school board, and teaching language (French or English).

The unit of observation in the QSCDK data is the student. From the overall sample, we exclude students in private schools, which are not subject to the class size regulation (4.6 percent), and students in mixed[8] classes (1.4 percent). We are left with 58,949 students in more than 3,600 classes spread across 1,484 schools. The average class size is 17.5 and the average enrollment is 59.0. Figure 1 shows the class size distribution in our sample. Class size varies between nine and 24,[9] with less than two percent of students being in classes of more than 21 students. Average student characteristics are reported in Table 1. Teachers filled out the questionnaire near the end of the school year, when students were six years old on average. Most students speak French at home (77%), attended childcare prior to kindergarten (62%), but did not attend pre-kindergarten in a school setting (79%). Around 6% of the children have learning disabilities or behavioral issues. Speech difficulties are present for 5% of the children. Indicators of health issues suggest that most children are in good health.

For ease of interpretation, we standardize each of the development measures that we

[7]Social and material deprivation indices are provided in the database. The Institut de la statistique du Québec computes the indices using a principal component analysis based on five measures of the 2006 Census data for each dissemination area (about 400 to 700 persons in Canada) (Section 2.3, Institut de la statistique du Québec, 2013b). Social deprivation mainly reflects a high proportion of individuals that are separated, divorced or widowed, and a high percentage of single families in the area. Material deprivation mainly reflects a high proportion of high school dropouts, a low proportion of individuals employed, and low average income in the area.

[8]Mixed classes contain students from different grade levels, for example, pre-kindergarten students and kindergarten students.

[9]Disclosure rules do not allow us to release the number of classrooms at size 24 because it is too small, but students in those classes are in the estimation sample.

use as outcomes to have a mean of zero and standard deviation of one,[10] with larger values meaning a more positive outcome. Figure 2 shows the average developmental scores by class size. It appears that averages of each of the four main scores are decreasing with class size between nine and 14 students, but beyond that point, average scores seem to be relatively stable, and even slightly increasing with class size. This pattern holds for all measures of child development but is strongest for cognitive development and communication skills. Figure 2 motivates us to estimate nonlinear models of the effect of class size on child development. Figure 2 also shows that the standard errors do not increase with class size, thus minimizing concerns about measurement error in the survey instrument coming from teachers in large groups reporting less accurately.[11] The next section will explain how we tackle identification issues and control for observables to obtain causal estimates.

5 Empirical strategy

Our goal is to estimate the effect of class size ($ClassSize_{ijs}$) on various measures of child development (y_{ijs}), where i is a child, j their class, and s the school, and to assess whether this effect is nonlinear. While class size may positively impact student test scores through reduced student disturbance as modeled by Lazear (2001), student ability may also influence student assignment to different class sizes. To identify the causal relationship between class size and student scores, we cannot solely rely on correlation between scores and class size; ordinary least squares (OLS) estimations may yield biased estimates. A first approach to pin down the causal effect of class size is to look at experimental evidence, such as the one provided by the influential Project STAR.[12] A second approach is to follow an instrumental variable

[10] Our results are also robust to using percentile ranks instead of standardized (or raw) scores.

[11] Larger standard errors in smaller classes mostly come from the fact that such classes are less common (smaller n).

[12] Project STAR, which took place between 1985 and 1989 in Tennessee, randomly assigned kindergarten to grade three students to regular or small classes (Finn and Achilles, 1999). Krueger (1999) and Krueger and Whitmore (2001) find that students assigned to smaller classes had higher test scores compared to students assigned to regular sizes. Black students and those receiving a free lunch benefited more from being assigned to a small class, and most of the effects documented were found for medium to high performing students (Konstantopoulos and Li, 2012; Jackson and Page, 2013; Mueller, 2013). In a study on the long-term effects of Project STAR, Chetty et al. (2011) find that students in smaller classes were more likely to attend college, but that teacher experience and peer quality also matter for long-term outcomes such as college attendance

(IV) approach, such as the one exploited by Angrist and Lavy (1999),[13] who explained that in Israel, enrollment and class size were positively related through Maimonides' rule, a rule dictating the maximum number of students per class. We follow this second approach: we use governmental regulation on class size that sets the maximum number of students per class per grade to identify the causal impact of class size on student achievement but depart from Angrist and Lavy (1999) in that we do not directly use the class size rule as an instrument. Our approach is closer to Fredriksson et al.'s (2013) fuzzy RD design, who use the different thresholds as points of discontinuity at which class size jumps, thus providing an instrument for class size.

In Québec, kindergarten class size cannot exceed 20 students. We start by splitting our sample in five different sub-data sets based on school enrollment segments: enrollment 10 to 40; 21 to 60; 41 to 80; 61 to 100; 81 to 120. As a result, each sub-data set is centered around a discontinuity threshold (21, 41, 61, 81, and 101). We then normalize enrollment so that all the thresholds are at 0, and stack the sub-data sets. When doing so, some observations are repeated because they can be in two windows, but we control for enrollment segments and cluster our standard errors at the segment level.

Our empirical approach relies on the following models:

$$y_{ijs} = \alpha + \pi Z_{ijs} + \gamma V_{ijs} + \delta Z_{ijs} \cdot V_{ijs} + \theta X_i + \psi W_s + \varepsilon_{ijs} \tag{1a}$$

$$ClassSize_{ijs} = \phi + \eta Z_{ijs} + \varphi V_{ijs} + \kappa Z_{ijs} \cdot V_{ijs} + \theta_1 X_i + \psi_1 W_s + \epsilon_{ijs}, \tag{1b}$$

where y_{ijs} is the outcome (development score) of student i in class room j and school s, and $ClassSize_{ijs}$ is their class size. We normalize enrollment to be centered at the threshold: V_{ijs} is the resulting normalized enrollment variable, where $V = 0$ is the threshold. Finally,

and earnings.

[13] Angrist and Lavy (1999), using data on third to fifth graders in Israel, exploit exogenous variations in class size due to school enrollment to instrument for class size. This approach has been applied to a number of different contexts: in Bolivia (Urquiola, 2006), Denmark (Browning and Heinesen, 2007; Nandrup, 2016), Japan (Akabayashi and Nakamura, 2014), the United States (Cho et al., 2012; Chingos, 2012), and Sweden (Fredriksson et al., 2013), to name a few. The estimated effects of class size reduction are often positive, but modest, and sometimes not statistically different from zero. Other approaches include using enrollment variations across time (Hoxby, 2000), or a general equilibrium framework (Gilraine et al., 2018).

we construct Z_{ijs} as a dummy variable equal to one when enrollment is to the right of the threshold ($Z = \mathbb{1}(V \geq 0)$). This specification allows slopes to be different on either side of the threshold. Equation 1b is the first stage, with coefficient η representing the effect of crossing the threshold on class size. Equation 1a is the reduced form. In practice, we estimate those two equations using two-stage least squares, which directly gives us the IV coefficient ($\beta = \pi/\eta$) and its standard error.

Our models also contain student- (X_i) and school-level (W_s) covariates. Student controls (X_i) include dummy variables indicating gender, student's age in months, place of birth, whether the child attended childcare or pre-kindergarten, ten markers of health and behavioral problems (physical disability, visual deficiency, auditive deficiency, speech disorder, learning difficulties, emotional problems, behavioral problems, disadvantaged family environment, chronic health conditions, and dental problems), and dummies to indicate whether the child received help from various school professionals (nurse, speech therapist, psychoeducator, social worker, and psychologist). School controls (W_s) include poverty index (high or low), social and material deprivation indices (highly advantaged, average, highly disadvantaged), teaching language (French or English), and school board dummies. Enrollment segments are used as controls. As mentioned above, the segments define bands of school enrollment: 10 to 40 students in the school; 21 to 60; 41 to 80; 61 to 100; and 81 to 120.

We can estimate Equations 1a and 1b using all sub-data sets pooled (i.e., all jumps) to get an average effect of class size on student development, or baseline effect $\bar{\beta}$. Results from this traditional linear approach are presented in Section 6.3. Angrist and Lavy's (1999) and Fredriksson et al.'s (2013) implicitly assume that the impact of class size is linear, or that the impact of class size is uniform across the class size distribution per unit of treatment. If the impact of class size reduction is not uniform, but instead varies across the class size distribution, then this strategy captures the average effect of class size reduction in the studied sample.

What makes our contribution innovative is that we modify this framework and exploit the discontinuities at the different jumps in class size as enrollment at the school level increases to test for non-linearity. In our application, going from a school enrollment of 20 to 21

induces a different jump in class size than going from an enrollment of 40 to 41, and so on. These jumps can be seen in Figure 3, which illustrates the predicted class size along with the actual average class size given enrollment. This figure clearly shows that the average class size follows tightly the government rule, albeit more so when the enrollment does not exceed 80 students.

To get at a nonlinear effect of class size, we compare estimates obtained at the first jump (β_1) to those obtained using the other jumps (β_{2+}).[14] We test for the equality of the two β coefficients and conclude that there is a nonlinear effect if we can reject the null of equal coefficients. In our main specification, we compare the jump in segment 1 (enrollment 10 to 40) to the jump in segments 2 and 3 together (enrollment 21 to 60 and 41 to 80, respectively). We do so because the bulk of our students are in schools with enrollment in those three segments (more than 75%), but we later assess the robustness of our estimates to the inclusion of more segments.

We also use, as another robustness check, a "donut" approach, where we remove the points right around the threshold. More specifically, we exclude students around the discontinuity points (the "donut hole"), i.e. in schools with an enrollment of 20, 40, and 60 (these are the schools observed just before the jump), and in schools with an enrollment of 21, 41, 42, 61, 62 and 63. In the agreement between the Québec government and school teachers, there is a clause[15] that allows deviations from the 20-students-per-class rule. Indeed, class size can exceed 20 under certain conditions, namely the lack of classrooms or qualified teachers in the region, but compensation to the teachers must be provided. In Figure 1 we saw that while deviations from 20 are limited, classes of 21 students are sometimes observed. The "donut" approach therefore excludes students in schools above the threshold but for which the average class size does not exceed 21.

The literature has highlighted a number of potential threats to the validity of the empirical approach based on school enrollment: score manipulation, bunching around the thresh-

[14] In practice, we do so by interacting the Z variable with the segments, thereby resulting in a model where the outcome is explained by two class sizes (segment 1 and segments 2+), instrumented by the two jumps.

[15] Annexe XVIII, Entente Nationale, Comité patronal de négociation pour les commissions scolaires francophones (CPNCF), Avril 2011)

olds, and class composition. We address each of these concerns as they come up in the next section. On class composition, while this is a legitimate concern in most applications, in our setting student assignment is done before most children ever set foot in the school. Since there are no formal modes of evaluation in childcare or pre-kindergarten in Québec, teachers have very limited knowledge of each student's ability when they assign students to different classes. In this sense, studying the impact of class size in kindergarten is particularly attractive since it is less likely that class size within a school depends on student ability. If students with more difficulties are assigned to smaller classes within the same school, then empirical evidence using OLS would detect a negative impact of class size on student achievement. However, in the methodology described above, enrollment relative to the threshold is used as opposed to actual class size, such that it is mainly variation in class size between schools that drives the identification, as opposed to variation within schools. Lastly, we cluster standard errors at the enrollment segment level in all of our estimations. Clustering at the school-district level leads to similar standard errors.

6 Findings

We now present the results from our empirical analysis. We start with a graphical description of our data and follow with tests on bunching and the balancing of covariates. Before we present estimates of our nonlinear empirical strategy, we start by showing results from linear models, both using ordinary least squares and IV. This allows us to compare our setting with previous studies. We then focus on our main results based on the nonlinear approach presented in the previous section. We also show some robustness analyses and investigate the heterogeneity of our nonlinear estimates. We conclude this section with a back-of-the-envelope estimation of the convexity of the education production function, linking back to the conceptual framework presented in Section 2.

6.1 Graphical description

Before we present findings from our model estimations, we start with a graphical description of the data. Figure 4 reproduces Figure 3, but this time showing on the Y-axis the average residual of class size after controlling for school-board fixed effects, in effect providing a visual version of the first stage when we exploit each enrollment threshold as an instrument (Equation 1b). The lines are fitted to individual-level data, by band of school enrollment. The drops at each vertical line (i.e. when school enrollment passes a threshold) are substantial and statistically significant. Since class size residuals become noisier above 80 and the number of observations becomes smaller, we focus on the first three segments, but test the sensitivity of our results to the inclusion of segments 4 and 5 in Section 6.5.

Figure 5 shows the corresponding jumps in cognitive development, by plotting the average residual of cognitive development by school enrollment, after controlling for student- and school-level covariates. At each threshold we observe an increase in the cognitive development residual. The fitted lines by band in Figure 5 also all slope downward, except for the 41-to-61 enrollment band, for which the line is rather flat. Those two figures, which in essence replicate Figures 5 and 6 in Fredriksson et al. (2016), show how the fuzzy RD strategy operates: the class size rule induces drops in class size, which are matched to jumps in developmental scores.

6.2 Diagnostics

Before delving into our empirical results, we present some balancing tests and investigate the possibility of bunching. In the present application, one might worry that the number of schools is not evenly distributed around the different enrollment thresholds, and that parental and student characteristics differ around the thresholds (Urquiola and Verhoogen, 2009; Fredriksson et al., 2013, 2016). Fredriksson et al. (2013, 2016) show evidence of bunching. They find that the fitted value of class size predicts parental education, such that parental education is not distributed uniformly around the thresholds.[16]

[16] In 1962, Sweden implemented a compulsory school law, which induced schools to change their catchment area in favor of those most in need. To address this issue, the authors use a one-school district approach to

In the absence of manipulation, the density of the observed units should be continuous around the threshold or alternatively, bunching on either side of the threshold should not be observed (McCrary, 2008). Class size does however exhibit some form of bunching since local school districts were not delimited randomly in the past, and were by design constructed to reach a certain fraction of the population. School catchment areas are rarely modified, at least as long as new schools are not added. To the best of our knowledge, no new school opened in the province in 2012. Furthermore, the public education law prevents school principals from refusing school access to a child living in the catchment area of the school board, so long as the school has not reached full capacity. In this sense, while some form of bunching should be expected if school catchment areas are well defined to start with, strategic bunching due to principals' behavior should not be observed.

Another form of bunching could however emerge because of parental choices. In Québec, each child is assigned to a school through the school's catchment area. However, parents may choose a different school within the catchment area of the entire school board they live in. Parents are however encouraged to choose their local school because their child could be displaced at the start of any school year if the chosen school reaches its capacity. In this case, priority is given to students within the school catchment area. In practice, it is therefore possible that schools of higher perceived quality are always closer to full capacity, and therefore have larger class sizes. Parents who opt out of their local school are likely to be parents who are more involved in the schooling of their child. This implies that we could underestimate the true impact of class size reduction.

We formally investigate potential bunching in our data. We start by testing for bunching around the enrollment thresholds using various bandwidths. Table 2 presents the test results, which were conducted on the full sample, and on restricted samples. When we follow Fredriksson et al. (2016) and use a bandwidth of five on the full sample, we find no evidence of bunching (p-value of 0.183). However, when we use alternative bandwidths (optimal bandwidth and values of six or seven), the result no longer holds, and the p-values are under 0.05. We then tested for bunching at the region and school board levels, and our investigation

study the impact of class size on long-term outcomes.

revealed that we could not rule out bunching in two regions, and more specifically five school boards [17]. When these two regions (or five school boards within these regions) are excluded, the tests (reported in the second and third panels of Table 2) no longer support the presence of bunching around the enrollment thresholds, except when a bandwidth of seven is used. Based on these findings, we later test the robustness of our main results to the exclusion of these school boards and find that our results are robust.

Another valid concern may be that students, teachers and/or school resources below and above the enrollment thresholds are different, so that we are in fact measuring not only the impact of class size, but the impact of a bundle of changes related to class size. For our fuzzy RD strategy to be valid, there should be no discontinuity in the student and school characteristics around the enrollment thresholds. For example, a concern might be that students with more difficulties are assigned to smaller classes or that teacher quality differ around the thresholds, with more inexperienced teachers being sent to schools forced to open a new classroom. In our data, we have information on students' disabilities and information on professionals in the school, but we do not know teacher quality or experience. However, from an administrative standpoint, the order in which teachers choose their school is based on experience. It is not clear that inexperience teachers would be more likely to end up in schools forced to open a new class, since more experienced teachers could also choose that school. Many factors enter a teacher's choice of school, but the number of students in a class is not likely to be at the top of their list since they make a choice for more than a year and also because the number of students in the class is not provided at the time of their choice. Finally, other resources, such as the number of psychoeducators, are unlikely to change around the thresholds since they depend on total enrollment. Nonetheless, here we formally check for this possibility.

In Table 3, we look at school and student characteristics around those thresholds, and we do not find evidence of jumps at the discontinuity. Column 1 shows the estimated coefficients of a regression of the cognitive development score on the main covariates in our data. Results

[17] Confidentiality regulations regarding data usage prevent us from naming these regions or investigating the reasons that might have led to stronger evidence of bunching.

show that these covariates are highly correlated with cognitive development (this is also true for the other development measures, but is not presented here for brevity). Given the relationship between these variables and our outcomes, we need to check that they balance around the discontinuity point. Columns 2 to 6 present the results of individual estimations of each of the covariates on threshold dummies (call them Z_1 to Z_5), meaning each line refers to a separate regression. The coefficients on threshold dummies are all virtually equal to zero. This suggests that the covariates are not related to the instruments (Z_1 to Z_5). We also test that all coefficients on threshold dummies are jointly equal to zero. The q-values[18] associated with these F-tests are reported in Column 7, and are above 0.1 for all covariates except one (whether the child receives help from a social worker, q-value = 0.08). It appears that more children above the second threshold (Z_2 equals one if enrollment is above 40) receive help from social workers. Schools with larger enrollment are more likely to have access to different professionals. Receiving help from a social worker is negatively correlated with cognitive development (Column 1). Since children are more likely to be in larger schools with larger classes, our class size estimate would be overestimated if we did not control for these differences, which we do. Overall, we conclude that our observed covariates are not correlated with the thresholds, which suggests that student and school characteristics are distributed evenly across the thresholds.

Our final balancing test is presented graphically in Figure 6. To produce this figure, we use our stacked data set and estimate a linear model of cognitive development scores on student- and school-level baseline covariates, and enrollment segment dummies. All five segments are included. We then predict the cognitive score using this model, plot the average predicted score by value of normalized enrollment, and fit two linear regressions, one on each side of the threshold. We do see a drop in the fitted lines at the threshold, but the difference of -0.015 is not statistically different from zero (p-value = 0.325).

Other inputs correlated with student performance and not observed in our data could still be unbalanced. Our results should therefore be interpreted with this caveat in mind.

[18]The q-values account for the fact that we are testing multiple hypotheses at the same time. We follow Simes (1986).

As mentioned above, teachers' experience is one of them. While we can't observe teachers' experience or quality, we can observe how each teacher answers the 104 questions on child development. The pattern of their answers may reveal inexperience or laziness. We come back to this possibility in the robustness check section.

6.3 Linear models

For the sake of comparison with the literature estimating linear effects, we first present estimates of naive linear models of class size on development scores. Those can be found in Columns 1 to 3 of Table 4. In Column 1, we estimate a linear model using ordinary least squares without controls. We add school-level controls in Column 2, and student-level controls in Column 3. Using OLS without controls, we find that both cognitive and noncognitive scores tend to increase with class size. Estimated coefficients are generally positive and statistically different from zero, but their magnitude is very small, ranging between 0.006 to 0.013 of a standard deviation (SD). These results differ from Hoxby (2000), who found negative and significant coefficients using the naive model. This difference provides some support to the idea that in contrast with Hoxby (2000), whose setting allowed parents with unobserved good characteristics to self-select their child into schools with smaller classes, here parents do not have that option. Once we include school-level controls[19], the relationship becomes economically and statistically equal to zero for all scores except emotional maturity (0.003 SD). Adding student-level characteristics[20] as controls further reduces the estimated coefficients on class size, but they remain not statistically different from zero, except for emotional maturity (0.002 SD). Overall, our naive OLS results accounting for student and school characteristics would suggest that class size has no effect on student development.

We then instrument class size using Z_{ijs} and its interaction with enrollment to estimate

[19]School controls include normalized enrollment, poverty index (high or low), social and material deprivation indices (highly advantaged, average, highly disadvantaged), teaching language (French or English), and school board dummies.

[20]Student controls include dummy variables indicating gender, student's age in months, place of birth, whether the child attended childcare or pre-kindergarten, ten markers of health and behavioral problems (physical disability, visual deficiency, auditive deficiency, speech disorder, learning difficulties, emotional problems, behavioral problems, disadvantaged family environment, chronic health conditions, and dental problems), and dummies to indicate whether the child received help from various school professionals (nurse, speech therapist, psychoeducator, social worker, and psychologist).

$\bar{\beta}$ (π/η). This approach is in line with that of Fredriksson et al.'s (2013). We find that increasing class size has an average negative impact on all of our outcomes, except for emotional maturity. For example, increasing class size by one student reduces cognitive development by 0.014 SD. This effect may appear small, but it needs to be compared to other estimates of class size effects in kindergarten. In the STAR experiment, kindergarten students who benefited from a small class size assignment scored higher on the Stanford Achievement Test[21] by an average of 0.20 SD (Krueger, 1999). Small classes had on average 15.1 students in kindergarten compared to regular groups of 22.4 students. Krueger's STAR estimate thus measures the impact of a variation of 7.3 students on average. If we multiply our estimates by 7.3, we get an impact of 0.102 SD for cognitive development, about half the size of Krueger's STAR. Our estimated impact is therefore about half of the impact documented in Krueger (1999). If we restrict our attention to children not around the discontinuity point, our "donut" approach (Column 5), we find slightly stronger results (0.019 SD for cognitive development), which brings us closer to the Krueger estimate.

The above linear models do not however capture nonlinearities. Using the log of class size could account for some nonlinearities but would not allow us to test whether one model is better than the other. Another possibility, under the assumption that class size is exogenous, is simply to leave aside the RD approach and regress developmental scores on class size dummies to lift the linear restriction imposed by a model in which a continuous class size variable explains development. Doing so (and including student- and school-level controls) yields estimates that present a nonlinear pattern shown in Figure 7. This pattern is similar to that observed in Figure 2. Endogeneity may still be an issue, but traditional statistics to test for endogeneity in an IV context are uninformative if the impact per unit of treatment is not uniform (Lochner and Moretti, 2015). Lochner and Moretti propose an exogeneity test in a context similar to ours, and show that if the test fails to reject exogeneity, then the OLS estimator is consistent and IV is not required. We thus perform the test proposed by Lochner and Moretti, and find that class size remains endogenous for most outcomes, even

[21] The Stanford Achievement Test measures achievement in reading, word recognition, and math in kindergarten.

in the context of varying treatment effects.[22] We therefore have to move to our nonlinear IV approach to address both endogeneity and allow for nonlinearity.

6.4 Nonlinear effects of class size

We may reasonably expect class size variations to have a larger impact in smaller classes (Lazear, 2001). One explanation may be that a student's marginal impact on the overall learning environment of the class is larger the smaller the number of students in the class. For example, talking with one's neighbor is more likely to disturb a class of 12 students than a class of 30 students since the probability that other students are also talking at the same time increases with class size.

When we account for the endogeneity of class size using a fuzzy RD approach (Equations 1a and 1b) and allow our model to capture the potential nonlinear effect of class size (by comparing the jump at the first threshold to that of the other thresholds), we find a different narrative. As mentioned above, to do this, we stack the data from segments 1 to 3 and add a segment fixed effect to our model to capture the changes within each of the segment. For now we focus on segments 1 to 3 since over 75 percent of our students are observed within these three segments. Later we test the robustness of our results to the inclusion of segments 4 and 5.

Table 5 provides IV estimates for each of our four main outcome variables, as well as our three additional outcomes (math in Column 2, reading and writing in Column 3, and overall development in Column 7). In Table 5, each column within a panel is a separate regression. Only the coefficients of interest are presented (β_1 and β_{2+}), but all regressions include segment fixed-effects, and student- and school-level controls.

Looking at cognitive development (Column 1), we see that the marginal impact of increasing class size by one student within the first segment (β_1) is negative and significant at -0.013 SD. This suggests that reducing class size by one student increases the cognitive score by 0.013 SD. When we look at the impact of increasing class size within segments 2 and 3 (β_{2+}), we instead find a positive coefficient (0.008 SD), which would suggest that

[22] Test results are available from authors upon request.

increasing class size is beneficial. This second result is however not robust to the exclusion of children around the discontinuity points (Panel 2). However, using this donut sample, the marginal impact of class size becomes even stronger within the first segment at -0.020 SD. This effect represents about 75 percent of the effect size documented in Krueger (1999).

Within the first segment, class size drops by about 7.9 students, from around 20 students per class to 12.1 students per class. Within segments 2 and 3, class size drops only by 3.3 students, from around 20 students per class to 16.7 students per class. Class size variations are therefore not happening in the same part of the class size distribution. When we test whether the two coefficients are statistically different from each other, we get a p-value under 0.0001 in both Panels A and B, suggesting that β_1 and β_{2+} are indeed different at the 1% level. This suggests that the marginal effect within the first segment is larger than the marginal effect within the other segments, supporting evidence of a nonlinear relationship between class size and cognitive development.

Our cognitive development measure captures abilities in math and in reading and writing. Looking at them separately (Columns 2 and 3), we see that the effect of class size is similar in both language skills and mathematics, and again the null hypothesis, H0:β_1=β_{2+}, is rejected.

Turning to noncognitive measures, we find that class size reduction could help enhance social competence (Panels 1 and 2) and that the effect is nonlinear in class size. Our social competence measure captures the child's social skills, self-confidence, sense of responsibility, autonomy, work habits, and respect for peers, adults, and rules and routines. In this sense, it is a measure that captures the ability of the child to behave respectfully and autonomously in a group.

The evidence are not as clear and robust for emotional maturity and communication skills. Class size effects are only apparent when we focus on students not in schools around the discontinuity points (the donut sample). Emotional maturity is a measure that captures prosocial behavior and mutual aid, fear and anxiety, aggressive behavior, hyperactivity and inattention, and the expression of emotions. Communication skills refer to the ability to communicate in a way that is understood and the ability to understand others.

Finally, if we aggregate our four main measures into a single score, the development

index (Column 7), we find that class size reduction has a nonlinear positive impact on student cognitive and noncognitive development. This result holds in both panels.

In summary, we find that class size reductions could raise cognitive development as well as social competence. The effects we document are nonlinear, a finding which has important implications for the optimal design of policy interventions related to class size.[23] Measuring the average impact using a linear model thus possibly hides important variations along the class size distribution that further research will hopefully further document. Of course, class size reduction is a costly policy whose benefits need to be compared with that of other interventions in education. We come back to this point in the conclusion.

6.5 Robustness analyses

The literature on class size using class size rules or threshold dummies as instruments highlights two other threats to instrument validity: bunching and score manipulation. In this subsection, we assess the robustness of our previous results by performing various analyses. First, we add data from segments 4 and 5 (Panels 2 and 3, Table 6). Second, we validate the robustness of our results to the exclusion of regions where bunching could not be excluded (Panels 4 and 5, Table 6). Third, we test the robustness of our results to the reweighted of observations to take into account patterns most closely related to score manipulation (Panels 6 and 7, Table 6). Generally, we find that our main results hold.

Including students in segments 4 and 5, i.e. those in schools with an overall enrollment of more than 80 students, does not change our main findings (see Panel 2 of Table 6): class size reduction is associated with better cognitive development and social competence. The magnitude of the coefficients is similar. Nonlinearity of the effects is not as clear (see p-values, but remains stable once we restrict our attention to the donut sample (Panel 3). Figure 4 shows that class size residuals become noisier above 80 and the number of observations

[23] Nonlinear effects can also be captured by comparing a linear model to a model based on a quadratic function. In a previous version of this paper, we ran several other models, including a quadratic model in class size, instrumented using the class size rule and its square. The quadratic term being positive and significant for cognitive development and social competence (and the linear term negative), this also provided some evidence of the decreasing negative marginal impact of increasing class size along the class size distribution. We also investigated the use of a linear-log specification.

becomes smaller.

Also, despite the evidence presented in Section 6.2 that bunching is unlikely to be a substantive issue, we estimate our IV model on the subsample of our data that excludes students registered in the five school boards where bunching may have occurred (Panels 4 and 5, Table 6). We find that our main results continue to hold even when these school boards are excluded.

Another threat to the validity of our approach is that of score manipulation, which can occur because teachers are deliberately cheating, but also because they are inexperienced or simply shirking. In order to account for the possibility of score manipulation by teachers, we implement a procedure based on Quintano et al. (2009), which reweights the observations based on a fuzzy K-means clustering approach.[24] This procedure identifies clusters most likely to have faced manipulation (outliers) and provides new weights for each observation. More weight is given to scores that are less likely to have been manipulated. If teachers in smaller classes (or larger classes) are more likely to manipulate scores, this procedure would provide a reasonable strategy to correct for that. In our application, we exploit the pattern of answers by teacher for each of the 104 subquestions. For example, a teacher answering "*yes*" to every single question for all students in their class would be given much less weight in this procedure. This lack of variance in a teacher's answers could be caused by, for example, laziness, inexperience or unprofessionalism. More details on the procedure are provided in the online Appendix.

Panels 6 and 7 of Table 6 show our estimated coefficients when observations are reweighted to account for the possibility of score manipulation. The impact of class size on all measures of development is very similar to that of the baseline specification. Our results are therefore not driven by observations of the outlier clusters. This suggests that, to the extent to which teachers' behavior is reflected in the pattern of answers, our results are not driven by different teachers being assigned to smaller or larger classes.

Taken together, our robustness analyses show that our main findings hold under several conditions. The impact of class size reduction on the cognitive development of children living

[24] Angrist et al. (2017) use this procedure and provide evidence of score manipulation in Italy.

in disadvantaged areas appears very robust.

6.6 Heterogeneity of effects

To assess heterogeneity, we investigate whether the estimated nonlinear impacts of class size differ between certain subgroups of students. First we look at gender differences (Columns 1 and 2 of Table 7), then at differences related to the socioeconomic status (SES) of the student (Columns 3 and 4).

Estimates for girls and boys suggest that for both genders, the effect of being in a larger class is negative and statistically significant for cognitive development, reading and writing skills and mathematical skills. The estimates for noncognitive measures are also similar across gender, but with standard errors too large to reject a null hypothesis of zero effect. The bottom panel of Table 7 shows that boys have a lower mean score on all measures, except math, relative to girls, but unfortunately do not seem to benefit more from class size reduction.

We now turn to results relative to the student socioeconomic status. The QSCDK provides the child's quintile of material deprivation based on the location of the child's residence. The area of residence is defined at a very fine geographical level, the 6-digit postal code of the child's residence (only 19 households on average live in one postal code in Canada). This allows us to look at students living in high-poverty areas separately from others. To do so, we compare those in the top quintile (Q5) of material deprivation with those of the other four quintiles (Q1-Q4)[25]. Estimates for children living in the poorest areas (Q5 of material deprivation) are presented in Column 3 of Table 7, while estimates for all other children are presented in Column 4 (Table 7, Q1-Q4).

We find important differences between children from poorer neighborhoods compared to all others, but given the size of our standard errors, the differences across columns are not always statistically significant. Children in high-poverty areas benefit from class size reduction in many dimensions of their development. The estimated coefficient (β_1) on all

[25]The first quintile (Q1) refers to children living in Census dissemination areas at the bottom 20% of material deprivation (low-poverty areas), while the fifth quintile (Q5) refers to children living in dissemination areas at the top 20% of material deprivation (high-poverty areas).

seven measures are all important in magnitude and statistically significant. Since children in low-SES areas have lower mean scores (bottom panel), this implies that class size reduction, from around 20 to 12 students, could help reduce inequalities between low-SES and higher-SES children by boosting the development of low-SES children. Finally, we also estimated the impact separately for boys and girls by material deprivation quintile.[26] We find that boys in high-poverty areas benefit about three times more than girls when we look at the impact on cognitive development, but given the relatively small sample sizes for these subgroups, we cannot reject equal effects for both genders.

6.7 Back to the conceptual framework

In the presentation of our conceptual framework (Section 2), we showed that the sign of the second derivative of the education production function f with respect to educational input T (i.e., the sign of $f''(ClassSize) + f'(ClassSize)(2/ClassSize)$) was of particular interest to try to determine whether nonlinearities exist. To connect the production function to our empirical estimation, suppose the $f(.)$ function is a quadratic.

Our data allow us to identify two key parameters: the jump in cognitive score (fuzzy RD estimate) at the first enrollment threshold (school enrollment 20 to 21), and the jump at the second enrollment threshold (stacked thresholds two and above). Let's call the first parameter β_{20-c_1}, where c_1 is the class size right of the first threshold, and the second β_{20-c_2}. More generally, we have $\beta_{20-c_i} = E(e|c_i) - E(e|20) = \beta(c_i - 20) + \gamma(c_i^2 - 20^2)$. This gives us a system of two equations and two unknowns (β and γ). The solution for β and γ is presented in the online Appendix. This allows us to determine whether the education production function is convex in T, which will be true if $f''(ClassSize) + f'(ClassSize)(2/ClassSize) = 6\gamma + (2\beta/ClassSize) > 0$.

In practice, we estimate the threshold dummies β_{20-c_1} and β_{20-c_2} on the cognitive development score, as well as values of c_1 and c_2 computed from estimations of the class size jumps at the thresholds. We find the value of the second derivative of our quadratic production function numerically and compute its confidence interval using the variance-covariance

[26] Estimates are available from authors upon request.

matrix of our reduced-form estimates.

We find that the sign of the second derivative is indeed positive for all values of class size above 11, hence over the near-complete support of class sizes observed in our data. However, for class sizes of 12 to 16 (13 in the donut sample), the standard errors are too large to be able to reject the null hypothesis of a linear relationship. This back-of-the-envelope calculation is coherent with a convex education production function (with respect to class size). We are limited by the fact that our calculation is based on two point estimates. Our findings are thus indicative of a nonlinear relationship, but future research could try to better identify the shape of the production function.

7 Conclusion

Our results bring novel insights to the impact of class size on child development. Using previously unexploited census-based survey data, in line with a number of studies, we confirm that, on average, class size reduction increases both cognitive and noncognitive scores for kindergarten children. Our results are comparable to those of Krueger (1999). In terms of magnitude, the linear effect we estimate on cognitive development using IV is roughly half to three quarters of what Krueger (1999) had found for the STAR experiment, when we scale up our estimate to reflect a similar change. The impact on the EDI cognitive component is not trivial since it has been showed to be a strong determinant of school achievement over and above the cognitive assessments and direct school readiness test such as the PPVT and the Number Knowledge Test (Forget-Dubois et al., 2007).

Our main contribution is to confirm the presence of nonlinearities in the effect of class size on measures of cognitive and noncognitive development. Our findings suggest that estimates from linear models producing average effects mask nonlinearities. With nonlinearities, it becomes possible to reallocate a fixed number of teachers between classes of different sizes and raise the average scores of children. More specifically, it is possible to reassign teachers such that the gains for children allocated to smaller classes are larger than the losses of those assigned to larger classes. Drastically decreasing class size in a small number of classes

and allow for very small increases in a large number of classes could be optimal. At the individual level, the losses would be very small, but the gains for those benefiting from smaller classes would be important. Where it becomes even more promising is when we look at the heterogeneous effects of class size. We observe that children living in low-SES areas have lower average test scores on all EDI dimensions and that the effect of class size varies by SES. We find that children living in low-SES areas benefit greatly from smaller classes in terms their cognitive and noncognitive development. Higher-SES children, on the other hand, do not gain as much from smaller classes (no significant effects on noncognitive development and much smaller effects on cognitive scores). Given that the benefits are mainly concentrated among students from disadvantaged backgrounds (especially disadvantaged boys), who represent less than 20% of our population, it would be possible to redesign class-size practices with minimal harm for the majority while helping those who need it most. And this could all be achieved at a constant cost.

This is consistent with the policy advice in Piketty and Valdenaire (2006). This cost-neutral promising policy could benefit disadvantaged students without harming other students, especially if services from other professionnals are maintained across schools. Since the number of teachers would be kept constant, the quality of teaching would also remain constant. Hence, the potentially adverse general-equilibrium effects of a universal class-size reduction policy could be avoided. If redistribution across schools is contentious, class size reduction could target specific neighborhoods. Perhaps the same result could also be achieved by reducing class size more drastically in the earlier grades, and crowding the later grades within a school, given that Heckman (2006) shows that early interventions tend to have larger impacts. Finally, class size could also be reduced in specific subjects in which more difficulties are observed.

In Québec, class size in kindergarten was established at 22 until 1999, but reduced to 20 in 2000 and to 19 in 2016. In our data, using the 2012 allocation of students across schools, we simulated the impact of increasing the maximum class size rule from 20 to 22 while maintaining enrollment per school constant. Our results suggest that average class size would increase from 17.1 to 17.9 and that 203 fewer teachers and classes would be needed.

This would free up 4.4 percent of kindergarten teachers. If applied to all seven grades of primary school, a fairly small average increase in class size could free up a large number of human resources to reallocate to children with greater needs. Class size rules are often governed by union bargaining contracts. The cost of reducing class size across the board is non trivial, and it is not at all clear that reducing class size on average is cost effective. But what our results show is that without even thinking about changing the number of students or the number of teachers, it may be possible to help children most in need by allowing class size to vary even more substantially across different groups.

Limits to our analysis come from the contemporaneous nature of the data: we look at the effect of kindergarten class size on kindergarten outcomes. While early outcomes matter and are crucial to the long-term development of a child, our study does not have the breadth of long-term studies such as Chetty et al. (2011) or Fredriksson et al. (2013). A follow-up survey, or linkages using administrative files, would clearly help shed light on long-term effects. Another limit is that we only observe teacher-reported outcomes. While all teachers in Québec are professionals with a 4-year bachelor degree in education and teacher reports, after observing the child for at least six months, are likely to offer good evaluations of children relative to their classmates, the degree of standardization across teachers may not be as good as it would be if we were to use standardized testing. On the positive side however, concerns about teaching-to-the-test or score manipulation are probably more limited in our context. Moreover, as stated earlier, the relevance and reliability of the Early Development Instrument, which is the basis of the development measures used in our analysis, has been proven in the literature on child development. Finally, teachers were provided time outside of the classroom during their working hours to ensure they answer the questionnaire carefully and they were also guaranteed that their report would never be used to compute statistics at the school level (or even school board level), which certainly helped reduce the likelihood that teachers manipulate scores to improve the ranking of their school.

In conclusion, we highlight the nonlinear nature of the relationship between class size and child development, both cognitive and noncognitive. We study the impact within the class size bounds that we observe, that is mainly between 10 to 22 students per class, and

show that allowing for nonlinearities of the class size effect is important, and could lead to different policy implications than estimates coming from a typical linear model. Future research should embrace the puzzle of nonlinearity in the resource-allocation problem at the heart of the economics of class size.

8 References

Akabayashi, Hideo, and Ryosuke Kakamura. 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.

Angrist, Joshua D., Erich Battistin, and Daniela Vuri. 2017. In a small moment: Class size and moral hazard in the Mezzogiorno. *American Economic Journal: Applied Economics* 9(4): 216-249.

Angrist, Joshua D., and Victor Lavy. 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics* 114(2): 533-575.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul. 2010. Heterogeneous class size effects: New evidence from a panel of university students. *The Economic Journal* 120(549): 1365-1398.

Bedard, Kelly, and Peter Kuhn. 2008. Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review* 27(3): 253-265.

Brinkman, Sally, Tess Gregory, John Harris, Bret Hart, Sally Blackmore, and Magdalena Janus. 2013. Associations between the early development instrument at age 5, and reading and numeracy skills at ages 8, 10 and 12: A prospective linked data study. *Child Indicators Research* 6(4): 695-708.

Browning, Martin, and Eskil Heinesen. 2007. Class size, teacher hours and educational attainment. *The Scandinavian Journal of Economics* 109(2): 415-438.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. Simple local polynomial density estimators. *Journal of the American Statistical Association* 115(531): 1449-1455.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. *Quarterly Journal of Economics* 126(4): 1593-1660.

Chingos, Matthew M. 2012. The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review*, 31(2012): 543-562.

Cho, Hyunkuk, Paul Glewwe, and Melissa Whitler. 2012. Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review* 31(3): 77-95.

Dee, Thomas S., and Martin R. West. 2011. The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis* 33(1): 23-46.

Ding, Weili, and Steven F. Lehrer. 2011. Experimental estimates of the impacts of class size on test scores: Robustness and heterogeneity. *Education Economics*, 19(3): 229-252.

Dunn, Lloyd M., and Leota M. Dunn. 1997. Peabody picture vocabulary test (3rd ed.). Circle Pines, MN: American Guidance Service.

Finn, Jeremy D., and Charles M. Achilles. 1999. Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis* 21(2): 97-109.

Forget-Dubois, Nadine, Jean-Pascal Lemelin, Michel Boivin, Ginette Dionne, Jean R. Séguin, Frank Vitaro, and Richard E. Tremblay. 2007. Predicting early school achievement with the EDI: A longitudinal population-based study. *Early Education and Development* 18(3): 405–426. Available at https://doi.org/10.1080/10409280701610796

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2013. Long-term effects of class size. *The Quarterly Journal of Economics*, 128(1): 249-285.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2016. Parental responses to public investments in children: Evidence from a maximum class size rule. *Journal of Human Resources* 51(4): 832-868.

Gilraine, Mike, Hugh Macartney, and Rob McMillan. 2018. Education reform in general equilibrium: Evidence from California's class size reduction. NBER Working Paper 24191, National Bureau of Economic Research, Cambridge, MA.

Government of Québec. 2012. Mieux accueillir et éduquer les enfants d'âge préscolaire,

une triple question d'accès, de qualité et de continuité des services. Conseil supérieur de l'éducation. ISBN : 978-2-550-65006-5

Hanushek, Eric A. 1999. The evidence on class size. In Mayer, S. E., & Peterson, P. E. (Eds.), *Earning & Learning : How Schools Matter*, 131-168. Washington, D.C.: Brookings Institution Press.

Hanushek, Eric A. 2002. Evidence, politics, and the class size debate. In Mishel, L. & Rothstein, R. (Eds.), *The Class Size Debate*, 37-66. Washington, D.C.: Economic Policy Institute.

Heckman, James J. 2006. Skill formation and the economics of investing in disadvantaged children. *Science* 312(5782): 1900-1902.

Hojo, Masakazu. 2013. Class-size effects in Japanese schools: A spline regression approach. *Economics Letters* 120(3): 583-587.

Hoxby, Caroline M. 2000. The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115(4): 1239-1285.

Institut de la statistique du Québec. 2013a. Québec Survey of Child Development in Kindergarten (QSCDK) 2012 data files. Accessible at the ISQ Data Access Centres (CADRISQ).

——. 2013b. Enquête québécoise sur le développement des enfants à la maternelle 2012, Portrait statistique pour le Québec et ses régions administratives, ISBN 978-2-550-68877-8

Jackson, Erika, and Marianne E. Page. 2013. Estimating the distributional effects of education reforms: A look at project STAR. *Economics of Education Review* 32(2013): 92-103.

Janus, Magdalena, Sally A. Brinkman, and Eric K. Duku. 2011. Validity and psychometric properties of the early development instrument in Canada, Australia, United States, and Jamaica. *Social Indicators Research* 103(2): 283-297.

Janus, Magdalena, and David R. Offord. 2007. Development and psychometric properties of the early development instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science* 39(1): 1-22.

Konstantopoulos, Spyros, and Wei Li. 2012. Modeling class size effects across the achieve-

ment distribution. *RISE – International Journal of Sociology of Education* 1(1): 5-26.

Krueger, Alan B. 1999. Experimental estimates of education production functions. *The Quarterly Journal of Economics* 114(2): 497-532.

Krueger, Alan B. 2002. Understanding the magnitude and effect of class size on student achievement. In Mishel, L. & Rothstein, R. (Eds.), *The Class Size Debate*, 7-35. Washington, D.C.: Economic Policy Institute.

Krueger, Alan B. 2003. Economic considerations and class size. *The Economic Journal* 113(485): F34-F63.

Krueger, Alan B., and Diane M. Whitmore. 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *The Economic Journal* 111(468): 1-28.

Laurin, Isabelle, Danielle Guay, Michel Fournier, Nathalie Bigras, and Anabel Solis. 2015. La fréquentation d'un service éducatif préscolaire: un facteur de protection pour le développement des enfants de familles à faible revenu? *Canadian Journal of Public Health* 106(7): eS14-eS20.

Laurin, Isabelle, Danielle Guay, Michel Fournier, Danielle Blanchard, and Nathalie Bigras. 2018. Quelle est l'association entre les caractéristiques résidentielles et du quartier et le développement de l'enfant à la maternelle? *Canadian Journal of Public Health* 109(1): 35-42.

Lazear, Edward P. 2001. Educational production. *Quarterly Journal of Economics* 116(3): 777-803.

Lazear, Edward P. 2006. Speeding, terrorism, and teaching to the test. *Quarterly Journal of Economics* 121(3): 1029-1061.

Lochner, Lance, and Enrico Moretti. 2015. Estimating and testing models with many treatment levels and limited instruments. *Review of Economics and Statistics* 97(2): 387-397.

McCrary, Justin. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2): 698-714.

Ministère de l'Éducation, de l'Enseignement supérieur et de la Recherche (MEESR) 2015,

Indicateurs de l'Éducation, Éducation préscolaire, enseignement primaire et secondaire, Édition 2014. Gouvernement du Québec. Available at http://www.education.gouv.qc.ca/ filead-min/site_web/documents/PSG/statistiques_info_decisionnelle/indicateurs_2014_fr.pdf

Mueller, Steffen. 2013. Teacher experience and the class size effect – Experimental evidence. *Journal of Public Economics* 98(2013): 44-52.

Nandrup, Anne Brink. 2016. Do class size effects differ across grades? *Education Economics* 24(1): 83-95.

OECD. 2017. Education at a glance 2017: OECD indicators, Éditions OCDE, Paris. Available at http://dx.doi.org/10.1787/eag-2017-en

Okamoto, Yukari, and Rafael Case. 1996. Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development* 61(1/2): 27–58.

Piketty, Thomas, and Mathieu Valdenaire. 2006. L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français. Estimations à partir du panel primaire 1997 et du panel secondaire 1995, Paris: Ministère de l'éducation nationale, 2006.

Quintano, Claudio, Rosalia Castellano, and Sergio Longobardi. 2009. A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of the outliers on assessment test scores. *Statistica & Applicazioni*, VII(2), 149-171.

Rockoff, Jonah, and Lesley J. Turner. 2010. Short-run impacts of Accountability on school quality. *American Economic Journal: Economic Policy* 2(4): 119-147.

Simard, Micha, Marie-Eve Tremblay, Amélie Lavoie, and Nathalie Audet. 2013. *Enquête québécoise sur le développement des enfants à la maternelle 2012*, Québec, Institut de la Statistique du Québec.

Simes, Robert J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73(3): 751-754.

Urquiola, Miguel, and Eric Verhoogen. 2009. Class-size caps, sorting, and the regression-discontinuity design. *The American Economic Review* 99(1): 179-215.

Urquiola, Miguel. 2006. Identifying class size effects in developing countries: Evidence

from rural Bolivia. *Review of Economics and Statistics* 88(1): 171-177.

Wechsler, David. 1989. Manual for the Wechsler Preschool and Primary Scale of Intelligence–Revised. San Antonio, TX: Psychological Corporation.

Zelazo, Philip David, Sophie Jacques, Jacob A. Burack, and Douglas Frye. 2002. The relation between theory of mind and rule use: Evidence from persons with autism-spectrum disorders. *Infant & Child Development* 11: 171–195.

9 Figures

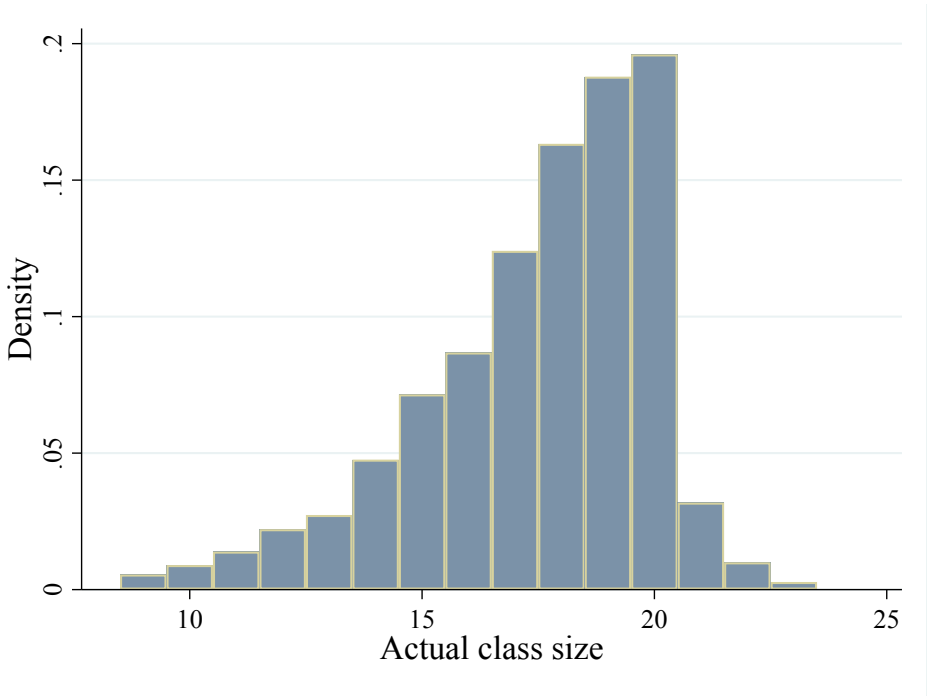

Figure 1: Histogram of Actual Class Size

Note: This figure shows the class size distribution in our data set, with class as a unit of observation.
Source: Authors' calculations using QSCDK 2012 data.

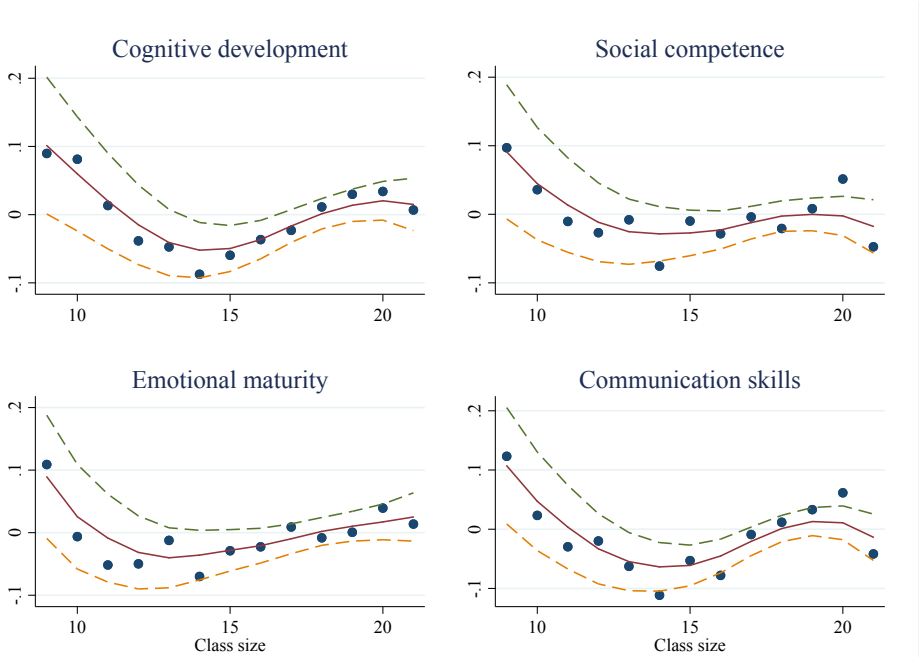

Figure 2: Average Developmental Score by Class Size

Note: In this figure, each dot corresponds to a class size (X-axis), and the Y-axis shows the average standardized developmental score for that class size. The solid line is a smoothed fit through the dots, while the dashed lines show 95% confidence intervals.
Source: Authors' calculations using QSCDK 2012 data.

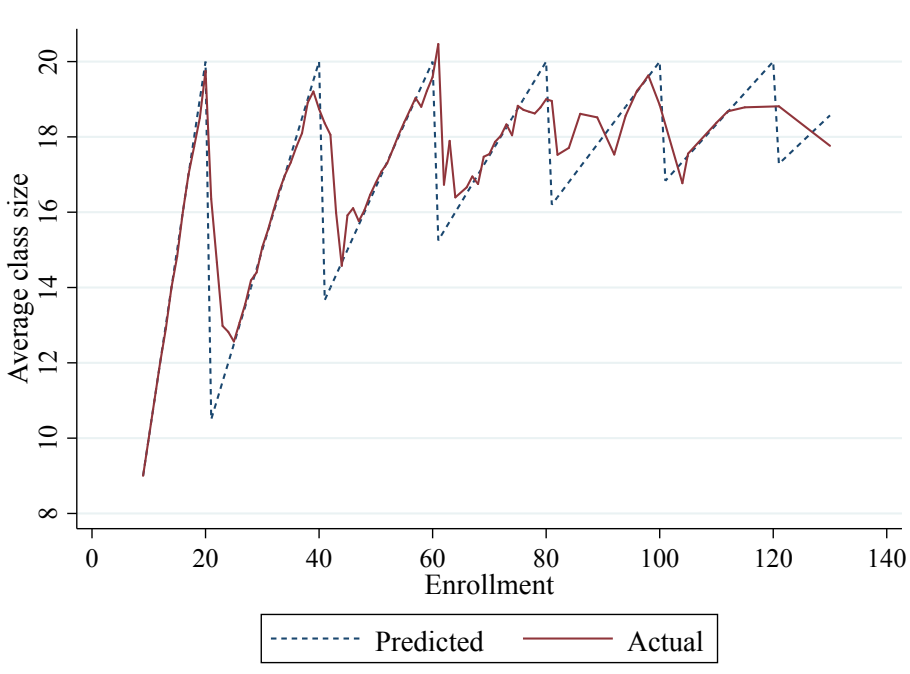

Figure 3: Average Predicted and Actual Class Size by School Enrollment

Note: This figure shows the actual (solid line) and predicted (dashed line) class size given enrollment.
Source: Authors' calculations using QSCDK 2012 data.

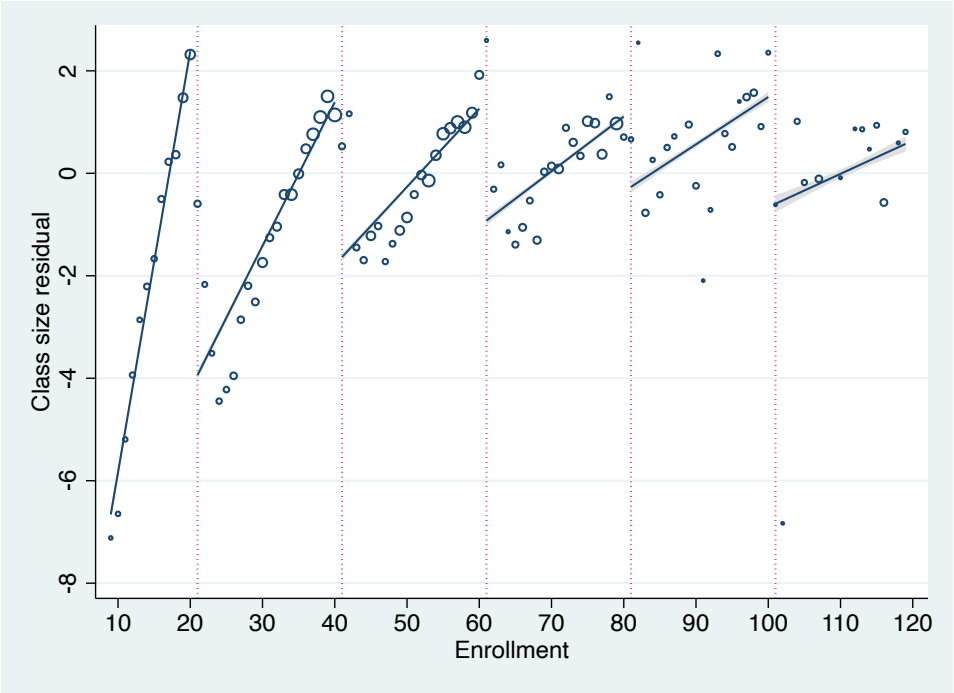

Figure 4: Class Size Residuals by Enrollment

Note: This figure shows residual class size, after controlling for school board fixed effects. The regression lines were fitted to individual data. The size of the dots is proportional to the number of observations.
Source: Authors' calculations using QSCDK 2012 data.

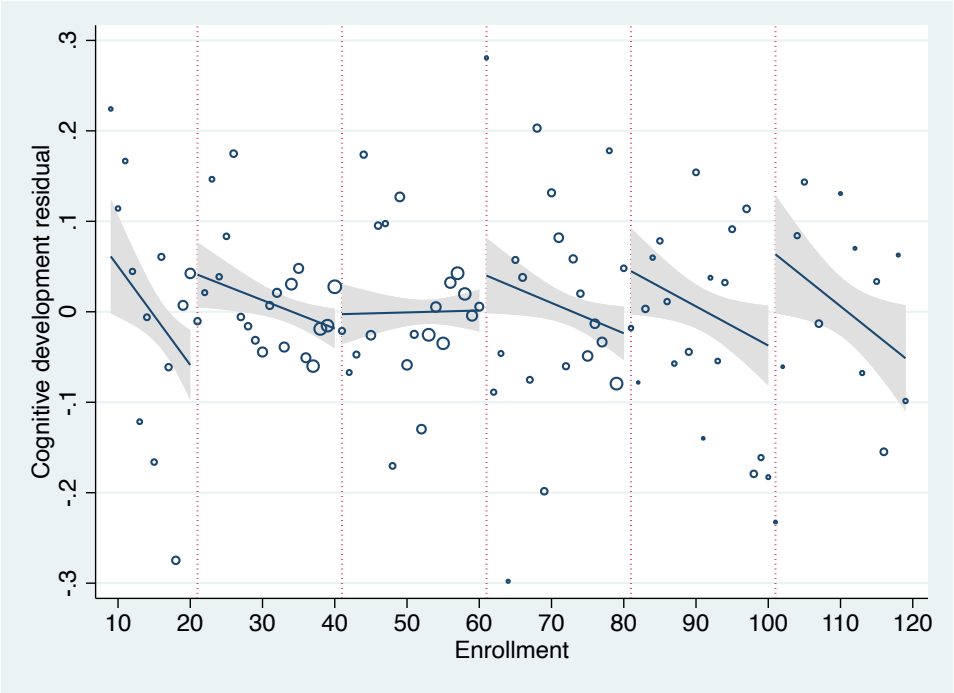

Figure 5: Cognitive Development Residuals by Enrollment

Note: This figure shows regression lines fitted to individual data of residual class size on enrollment, after controlling for student- and school-level covariates (see list in Table 4's note). The size of the dots is proportional to the number of observations.
Source: Authors' calculations using QSCDK 2012 data.

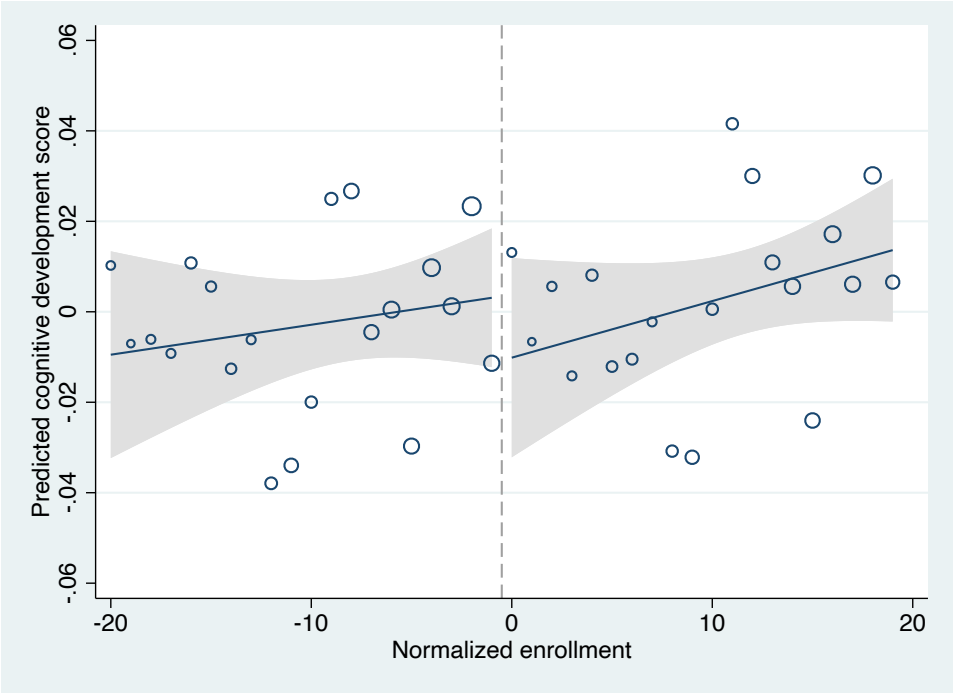

Figure 6: Balancing Test

Note: This figure plots the average predicted cognitive development score by normalized enrollment value, where the prediction comes from a linear model estimated by OLS and using student- and school-level covariates. Linear fits and their corresponding 95% confidence intervals are also shown, separately below and above the threshold (normalized enrollment value of 0). The size of the dots is proportional to the number of observations.
Source: Authors' calculations using QSCDK 2012 data.

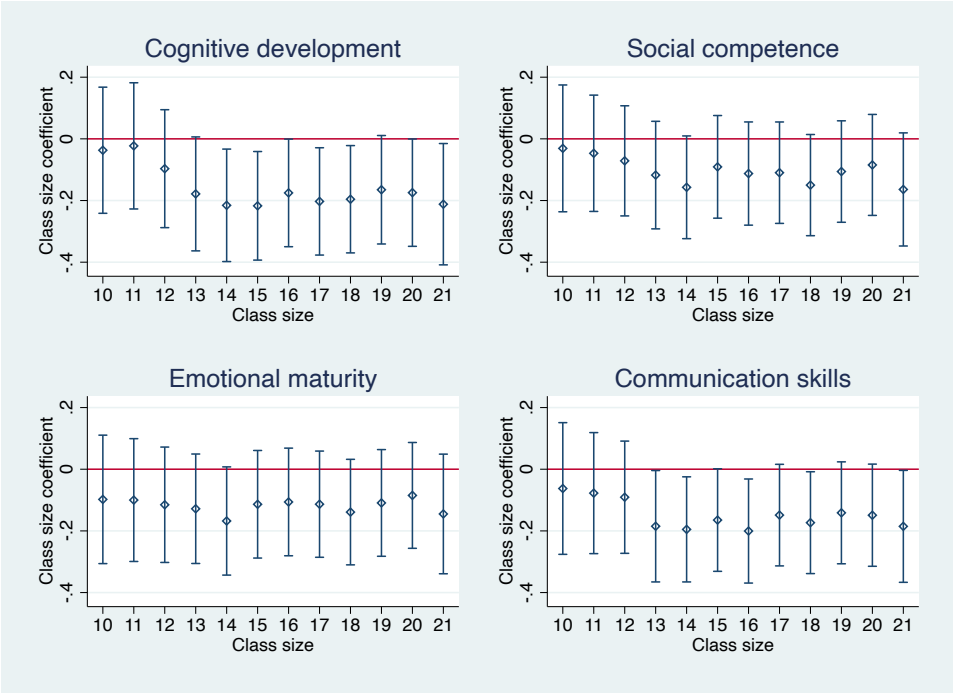

Figure 7: Nonlinearity: OLS Coefficients on Class Size Dummies

Note: This figure shows the estimated coefficients of class size dummies on developmental scores, including student- and school-level controls. Class size of nine is the omitted category. Bars show 95% confidence intervals, computed using standard errors clustered at the school level.
Source: Authors' calculations using QSCDK 2012 data.

10 Tables

Table 1: Descriptive Statistics - Student

	Mean
Age (in months)	72.07
Std. dev.	(3.62)
Female	0.50
Mother tongue	
French	0.77
English	0.07
Other	0.16
Dysfunctional family	0.030
Disabilities	
Physical	0.005
Dental	0.004
Visual	0.003
Hearing	0.004
Chronic disease	0.004
Speech	0.057
Learning	0.065
Emotional	0.031
Behavioral	0.063
Receiving help from a	
Nurse	0.03
Speech therapist	0.08
Psychoeducator	0.04
Social worker	0.02
Psychologist	0.03
Attended childcare	
Yes	0.62
No	0.15
Missing	0.23
Attended pre-kindergarten	
Yes	0.17
No	0.80
Missing	0.03

Note: $N = 58{,}949$

Source: Authors' calculations using QSCDK 2012 data.

Table 2: Density tests around enrollment thresholds

	Bandwidth		Effective N		p-value/	Excluding	Excluding
	Left	Right	Left	Right	t-stat	2 regions	5 school boards
RD density, p-value	5.20	4.33	567	188	0.044	no	no
(Cattaneo et al., 2020)	5	5	567	233	0.183	no	no
	6	6	664	274	0.031	no	no
	7	7	757	308	0.002	no	no
	5.75	4.25	397	153	0.382	yes	no
	5	5	397	186	0.721	yes	no
	6	6	465	220	0.314	yes	no
	7	7	542	246	0.091	yes	no
	5.01	4.24	494	177	0.215	no	yes
	5	5	494	219	0.277	no	yes
	6	6	582	256	0.096	no	yes
	7	7	673	289	0.017	no	yes
DC Density, t-stat					−0.138	no	no
(McCrary, 2008)					0.842	yes	no
					0.993	no	yes

Note: This table shows various test statistics testing for bunching around enrollment thresholds. The top panel shows p-values associated with Cattaneo et al.'s (2020) test, using the optimal bandwidth (first row), and bandwidths of five, six, and seven. Effective N refers to the number of schools within the bandwidths around the thresholds. The bottom panel shows the t-statistic from McCrary's (2008) test. Two regions or five school boards showing evidence of bunching are excluded sequentially from the sample, as indicated in the last two columns.
Source: Authors' calculations using QSCDK 2012 data.

Table 3: Children and school characteristics around thresholds

	This column is one regression	*Each row contains the estimates from a separate regression*					
	Cognitive development	Enrollment thresholds					q-value of F-test
		Z_1	Z_2	Z_3	Z_4	Z_5	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female student	0.150***	-0.014	-0.032	0.232	-0.262	0.046	0.683
	(0.009)	[0.970]	[1.000]	[0.811]	[0.958]	[1.000]	
Born in Canada	0.047**	0.018	0.017	-0.364	0.084	-0.009	0.523
	(0.023)	[1.000]	[1.000]	[0.965]	[1.000]	[0.993]	
Dysfunctional family	-0.866***	0.008	0.007	-0.092	-0.005	0.098	0.635
	(0.041)	[0.985]	[1.000]	[0.865]	[1.000]	[0.954]	
Disabilities							
Physical	-0.755***	-0.003	-0.024	0.012	0.023	-0.009	0.456
	(0.104)	[1.000]	[0.706]	[0.969]	[0.949]	[1.000]	
Visual	-0.381***	0.002	-0.014	-0.011	-0.031	0.063	0.380
	(0.115)	[1.000]	[0.973]	[0.969]	[1.000]	[0.536]	
Hearing	-0.681***	0.004	-0.005	-0.005	-0.004	0.007	0.725
	(0.092)	[0.425]	[1.000]	[1.000]	[1.000]	[1.000]	
Chronic disease	-0.371***	0.000	-0.008	-0.001	-0.023	0.038	0.817
	(0.110)	[1.000]	[0.976]	[1.000]	[0.997]	[0.874]	
Dental issues	-0.636***	0.002	0.004	0.032	-0.043	0.010	0.707
	(0.126)	[0.991]	[1.000]	[0.905]	[0.962]	[1.000]	
Help from a school professional							
Nurse	-0.180***	-0.079	0.015	0.246	0.014	-0.016	0.147
	(0.042)	[0.058]	[1.000]	[0.836]	[1.000]	[1.000]	
Speech therapist	0.564***	-0.059	0.124	-0.065	0.178	-0.049	0.198
	(0.028)	[0.052]	[0.806]	[1.000]	[0.988]	[1.000]	
Psychoeducator	0.067*	-0.022	0.092	0.049	-0.112	-0.012	0.668
	(0.035)	[0.698]	[0.938]	[1.000]	[0.975]	[1.000]	
Social worker	-0.146***	-0.031	0.169	-0.074	-0.116	0.138	0.082
	(0.048)	[0.120]	[0.063]	[0.959]	[0.988]	[0.979]	
Psychologist	0.290***	-0.026	0.061	0.005	0.085	-0.101	0.404
	(0.041)	[0.091]	[0.828]	[1.000]	[0.950]	[0.952]	
Other professional	0.355***	-0.06	0.084	-0.292	0.216	0.203	0.699
	(0.021)	[0.653]	[0.999]	[0.956]	[1.000]	[1.000]	
Attended pre-kindergarten	0.054***	0.138	-0.306	-0.013	0.277	-0.608	0.137
	(0.019)	[0.062]	[0.958]	[0.989]	[0.994]	[0.996]	
School poverty index (High)	-0.086***	-0.050	-0.176	-0.777	1.555	0.183	0.640
	(0.029)	[1.000]	[1.000]	[0.954]	[0.950]	[1.000]	
School language (French)	-0.171***	-0.059	-0.658	0.681	-0.656	1.025	0.454
	(0.043)	[0.842]	[0.661]	[0.820]	[1.000]	[0.962]	
N	53,760						

Note: Excludes school boards in which bunching cannot be rejected (N = 53,760). Column 1 shows the coefficients estimated using OLS for a model where the cognitive development score is explained by all the covariates (dummy variables) listed in the rows. Columns 2 to 6 report results from separate regressions (one per row) estimated by OLS, where the variable listed on the left is explained by the five thresholds (Z_1 to Z_5). Column 7 shows the q-value associated with the F-statistic testing for the joint significance of the coefficients on the five threshold dummies. Cluster-robust standard errors are reported in parentheses for specification (1) and q-values are reported in brackets under each Z_1 to Z_5 coefficient. In Column 1, significance is denoted using asterisks: *** is $p < 0.01$, ** is $p < 0.05$, and * is $p < 0.1$.
Source: Authors' calculations using QSCDK 2012 data.

Table 4: Linear class size effects

	OLS estimates			IV estimates	
	(1)	(2)	(3)	(4)	(5)
Cognitive development (N = 107,848)					
Class size	0.009**	0.001	-0.003	-0.014***	-0.016***
	(0.003)	(0.005)	(0.004)	(0.003)	(0.003)
Math (N = 99,291)					
Class size	0.008*	0.001	-0.002	-0.011****	-0.017***
	(0.004)	(0.005)	(0.005)	(0.003)	(0.003)
Reading & writing (N = 96,224)					
Class size	0.006**	-0.002	-0.006	-0.017***	-0.016***
	(0.002)	(0.004)	(0.004)	(0.003)	(0.003)
Social competence (N = 108,151)					
Class size	0.006	0.001	0.000	-0.002	-0.007**
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Emotional maturity (N = 107,401)					
Class size	0.008***	0.003**	0.002*	-0.005*	-0.012***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
Communication skills (N = 108,127)					
Class size	0.013***	0.007	-0.001	-0.007**	-0.008***
	(0.003)	(0.005)	(0.004)	(0.003)	(0.003)
Development index (N = 107,074)					
Class size	0.011**	0.004	-0.001	-0.009***	-0.013***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
School controls	no	yes	yes	yes	yes
Student controls	no	no	yes	yes	yes
Donut sample	no	no	no	no	yes

Note: Each cell contains the estimate from a separate regression using the stacked data from segments 1 to 5. Each regression includes segment dummies. School controls include normalized enrollment, poverty index (high or low), social and material deprivation indices (highly advantaged, average, highly disadvantaged), teaching language (French or English), and school board dummies. Student controls include dummy variables indicating gender, student's age in months, place of birth, whether the child attended childcare or pre-kindergarten, ten markers of health and behavioral problems (physical disability, visual deficiency, auditive deficiency, speech disorder, learning difficulties, emotional problems, behavioral problems, disadvantaged family environment, chronic health conditions, and dental problems), and dummies to indicate whether the child received help from various school professionals (nurse, speech therapist, psychoeducator, social worker, and psychologist). In this donut sample (Column 5), students in schools with enrollment between seg*20 to seg*21, with seg taking values between 1 and 5, are excluded. N in the donut sample is lower, going from 83,514 for the reading score to 93,838 for the social competence score. Cluster-robust standard errors are reported in parentheses. Significance is denoted using asterisks: *** is $p < 0.01$, ** is $p < 0.05$, and * is $p < 0.1$.
Source: Authors' calculations using QSCDK 2012 data.

Table 5: Nonlinear class size effects

	Cognitive development (1)	Math (2)	Reading & writing (3)	Social competence (4)	Emotional maturity (5)	Comm. skills (6)	Dev. index (7)
Panel 1: Segments 1 to 3							
Class size (β_1)	-0.013***	-0.017***	-0.014***	-0.006*	0.002	-0.003	-0.006**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	0.008**	-0.005	0.007*	0.003	-0.003	0.002	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
N	80,877	74,810	72,572	81,089	80,545	81,071	80,315
Test (p-value)	0.0000	0.0068	0.0000	0.0309	0.2154	0.2344	0.0351
Panel 2: Segments 1 to 3, donut sample							
Class size (β_1)	-0.020***	-0.026***	-0.018***	-0.015***	-0.008**	-0.009***	-0.016***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	0.000	-0.011***	0.002	-0.001	-0.008**	-0.001	-0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
N	71,636	66,358	64,321	71,811	71,361	71,793	71,168
Test (p-value)	0.0000	0.0015	0.0000	0.0008	0.9912	0.0366	0.0011

Note: Estimation based on stacked data from segments 1 to 3. Each column within a panel contains the estimates from a separate regression where the continuous class size variables (one for segment 1, one for segments 2+) are instrumented using segment-specific threshold dummies. Panel 1 is our benchmark specification. Panel 2 includes segments 1 to 3 but excludes students enrolled in schools around the discontinuity points. More specifically, in this donut sample, students in schools with enrollment between seg*20 to seg*21, with seg taking values between 1 and 3, are excluded. All models control for segment dummies, as well as student- and school-level characteristics. Each panel's bottom row reports the p-value of a test where $H0 : \beta_1 = \beta_{2+}$. Cluster-robust standard errors are reported in parentheses. Significance is denoted using asterisks: *** is $p < 0.01$, ** is $p < 0.05$, and * is $p < 0.1$.
Source: Authors' calculations using QSCDK 2012 data.

Table 6: Robustness checks

	Cognitive development (1)	Math (2)	Reading & writing (3)	Social competence (4)	Emotional maturity (5)	Comm. skills (6)	Dev. index (7)
Panel 1: Segments 1 to 3 (Benchmark, N = 80,877)							
Class size (β_1)	-0.013***	-0.017***	-0.014***	-0.006*	0.002	-0.003	-0.006**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	0.008**	-0.005	0.007*	0.003	-0.003	0.002	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Test (p-value)	0.0000	0.0068	0.0000	0.0309	0.2154	0.2344	0.0351
Panel 2: Segments 1 to 5 (N = 107,848)							
Class size (β_1)	-0.016***	-0.017***	-0.017***	-0.007**	0.000	-0.004	-0.009***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	-0.011***	-0.007*	-0.015***	0.001	-0.009**	-0.007*	-0.009**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Test (p-value)	0.2768	0.0433	0.6867	0.0524	0.0345	0.5881	0.9997
Panel 3: Segments 1 to 5, donut sample (N = 93,581)							
Class size (β_1)	-0.023***	-0.025***	-0.021***	-0.016***	-0.009***	-0.011***	-0.018***
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	-0.013***	-0.013***	-0.013***	-0.002	-0.013***	-0.006	-0.010***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Test (p-value)	0.0295	0.0117	0.1112	0.0010	0.4313	0.2214	0.0611
Panel 4: Segments 1 to 3, excluding bunching boards (N = 74,499)							
Class size (β_1)	-0.010***	-0.015***	-0.011***	-0.005	-0.001	-0.002	-0.006*
	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	0.013***	-0.001	0.013***	0.005	-0.006	0.002	0.004
	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Test (p-value)	0.0000	0.0044	0.0000	0.0336	0.2392	0.3193	0.0287
Panel 5: Segments 1 to 3, excluding bunching boards, donut sample (N = 66,280)							
Class size (β_1)	-0.021***	-0.026***	-0.018***	-0.016***	-0.010***	-0.010***	-0.017***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)
Class size (β_{2+})	0.003	-0.009**	0.004	-0.000	-0.009**	-0.001	-0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Test (p-value)	0.0000	0.0008	0.0000	0.0005	0.8293	0.0499	0.0006
Panel 6: Segments 1 to 3, reweighted for score manipulation (N = 80,877)							
Class size (β_1)	-0.011***	-0.014***	-0.012***	-0.007**	0.001	-0.003	-0.005
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	0.010**	-0.003	0.012***	0.005	-0.003	0.003	0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Test (p-value)	0.0000	0.0148	0.0000	0.0050	0.3279	0.1500	0.0266
Panel 7: Segments 1 to 3, reweighted for score manipulation, donut sample (N = 71,636)							
Class size (β_1)	-0.017***	-0.021***	-0.016***	-0.015***	-0.010***	-0.007***	-0.013***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Class size (β_{2+})	0.002	-0.010***	0.005	-0.002	-0.009**	-0.002	-0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Test (p-value)	0.0000	0.0096	0.0000	0.0013	0.7952	0.1070	0.0038

Note: Estimation based on stacked data from segments 1 to 3 or 1 to 5. Each column within a panel contains the estimates from a separate regression where the continuous class size variables (one for segment 1, one for segments 2+) are instrumented using segment-specific threshold dummies. Panel 1 is our benchmark specification. Panel 2 includes segments 1 to 5. Panel 3 includes segments 1 to 5 but excludes students in schools around the discontinuity points. Panel 4 includes segments 1 to 3 but excludes school boards in which bunching was detected. Panel 5 additionally excludes students in schools around the discontinuity points. Panel 6 reweights observations to account for the possibility of score manipulation or cheating, and panel 7 further excludes students around the discontinuity points. All specifications include segment dummies, school- and student-level controls. Each panel's bottom row reports the p-value of a test where $H0 : \beta_1 = \beta_{2+}$. The N presented correspond to the sample sizes in Column 1 (cognitive development); N varies slightly by column (see Table 5 for the benchmark model). Cluster-robust standard errors are reported in parentheses. Significance is denoted using asterisks: *** is $p < 0.01$, ** is $p < 0.05$, and * is $p < 0.1$.
Source: Authors' calculations using QSCDK 2012 data.

Table 7: Heterogeneous effects by subgroups

	Gender		Material deprivation	
	Girls	Boys	High (Q5)	Low (Q1-Q4)
	(1)	(2)	(3)	(4)
Cognitive development				
Class size (β_1)	-0.013***	-0.013***	-0.028***	-0.007**
	(0.004)	(0.005)	(0.008)	(0.004)
Class size (β_{2+})	0.009*	0.007	-0.018	0.014***
	(0.005)	(0.006)	(0.012)	(0.004)
Math				
Class size (β_1)	-0.016***	-0.017***	-0.030***	-0.011***
	(0.005)	(0.005)	(0.008)	(0.004)
Class size (β_{2+})	-0.001	-0.006	-0.020	0.000
	(0.006)	(0.006)	(0.013)	(0.005)
Reading & writing				
Class size (β_1)	-0.015***	-0.013**	-0.029***	-0.011***
	(0.004)	(0.005)	(0.008)	(0.004)
Class size (β_{2+})	0.009	0.006	-0.025**	0.012**
	(0.006)	(0.006)	(0.012)	(0.005)
Social competence				
Class size (β_1)	-0.005	-0.006	-0.027***	0.000
	(0.004)	(0.005)	(0.007)	(0.003)
Class size (β_{2+})	-0.002	0.008	-0.037***	0.011***
	(0.005)	(0.006)	(0.011)	(0.004)
Emotional maturity				
Class size (β_1)	0.002	0.002	-0.015**	0.007*
	(0.004)	(0.005)	(0.007)	(0.004)
Class size (β_{2+})	-0.007	0.001	-0.034***	0.003
	(0.005)	(0.006)	(0.011)	(0.004)
Communication skills				
Class size (β_1)	-0.001	-0.004	-0.023***	0.005
	(0.004)	(0.004)	(0.007)	(0.003)
Class size (β_{2+})	0.001	0.004	-0.020*	0.008**
	(0.005)	(0.006)	(0.011)	(0.004)
Development index				
Class size (β_1)	-0.006	-0.006	-0.028***	0.001
	(0.004)	(0.005)	(0.007)	(0.003)
Class size (β_{2+})	-0.001	0.006	-0.034***	0.010***
	(0.005)	(0.006)	(0.011)	(0.004)
N	39,992	40,885	14,102	64,814
Mean score				
Cognitive development	0.111	-0.113	-0.147	0.028
Math	-0.004	-0.003	-0.133	0.023
Reading & writing	0.154	-0.159	-0.140	0.025
Social competence	0.219	-0.219	-0.082	0.014
Emotional maturity	0.278	-0.279	-0.073	0.012
Communication skills	0.140	-0.123	-0.125	0.035
Development index	0.228	-0.227	-0.133	0.026

Note: Estimation based on stacked data from segment 1 to 3. Each column within a panel contains the estimates from a separate regression where the continuous class size variables (one for segment 1, one for segments 2+) are instrumented using segment-specific threshold dummies. All specifications include segment dummies, school- and student-level controls. The N presented correspond to the sample sizes for the cognitive development score; N varies slightly by outcome (see Table 5 for the benchmark model). Cluster-robust standard errors are reported in parentheses. Significance is denoted using asterisks: *** is $p < 0.01$, ** is $p < 0.05$, and * is $p < 0.1$.
Source: Authors' calculations using QSCDK 2012 data.