LONG-TERM EFFECTS OF CLASS SIZE*

PETER FREDRIKSSON
BJÖRN ÖCKERT
HESSEL OOSTERBEEK

This article evaluates the long-term effects of class size in primary school. We use rich data from Sweden and exploit variation in class size created by a maximum class size rule. Smaller classes in the last three years of primary school (age 10 to 13) are beneficial for cognitive and noncognitive ability at age 13, and improve achievement at age 16. Most important, we find that smaller classes have positive effects on completed education, wages, and earnings at age 27 to 42. The estimated wage effect is large enough to pass a cost-benefit test. *JEL* Codes: I21, I28, J24, C31.

I. INTRODUCTION

This article evaluates the effects of class size in primary school on long-term outcomes, including completed education, earnings and wages at age 27–42. While there is a large literature estimating the short-term effects of class size, credible estimates of long-term effects of class size are sparse.[1] To judge the effectiveness of class size reductions, it is vital to know whether short-term effects on cognitive skills (if any) persist or fade out, and whether these effects translate into economically meaningful improvements in labor market outcomes.

A few previous studies examine long-term effects of class size. The studies most relevant for us are based on the Tennessee STAR experiment. In STAR, students and their teachers were randomly assigned (within school) to different classrooms in grades K–3. Some students were randomly assigned to a

*We gratefully acknowledge comments from Lawrence Katz, Alan Krueger, Edwin Leuven, Per Pettersson Lidbom, Mikael Lindahl, Erik Lindqvist, Magne Mogstad, Helena Svaleryd, Miguel Urquiola, four anonymous referees, seminar participants in London, Mannheim, Paris, Stockholm, and Uppsala, and participants at various conferences. We acknowledge the financial support from the Marcus and Amalia Wallenberg Foundation and Handelsbanken.

1. Findings of short-term effects vary across countries, by age of the pupils and by empirical approach. Most studies that focus on class size in primary school and use a credible empirical strategy find that class size has a negative effect on cognitive achievement measured shortly after exposure. Well-known studies showing such effects are Angrist and Lavy (1999) for Israel, Krueger (1999) for the United States, and Urquiola (2006) for Bolivia. An equally well-known study finding no impact on U.S. data is Hoxby (2000).

The Quarterly Journal of Economics (2013), 249–285. doi:10.1093/qje/qjs048.
Advance Access publication on November 18, 2012.

class of around 15 students while others were assigned to a class of around 22 students. Krueger and Whitmore (2001) use this information and find that attendance of a small class in grades K–3 increases the likelihood of taking college entrance exams.

Until recently it was not possible to directly analyze the earnings impact of class size reductions using STAR data. Chetty et al. (2011) are, however, able to link the original STAR data to administrative data from tax returns. They find that students in small classes are significantly more likely to attend college and exhibit improvements on other outcomes. However, smaller classes do not have a significant effect on earnings at age 27. The point estimate is small, negative, and imprecise. The upper bound of the 95% confidence interval is an earnings gain of 3.4%. Following Krueger (2003) and Schanzenbach (2007), the authors compare this with a prediction of the expected earnings gain based on the estimated impact of small classes on test scores and the cross-sectional correlation between test scores and earnings. This imputed earnings estimate implies a positive effect of 2.7%, which—as the authors stress—lies within the 95% confidence interval of the directly estimated impact of small classes on earnings.[2]

There is also a literature on school quality and labor market outcomes. The most well-known paper is probably Card and Krueger (1992), who use U.S. data and exploit variation across cohorts within regions of birth to estimate the effects of measures of school quality on the returns to schooling and earnings. They conclude that a lower pupil/teacher ratio increases the rate of return to schooling and earnings (the latter conclusion has been criticized by Heckman, Layne-Farrar, and Todd 1996). Dearden, Ferri, and Meghir (2002) use a conditional independence assumption to estimate the relationship between school quality and earnings using U.K. data. The association between the pupil/teacher ratio and earnings is typically insignificant. Dustmann, Rajah, and van Soest (2003) use the same data and approach as Dearden, Ferri, and Meghir (2002) but assume that class size has no direct effect on wages conditional on educational attainment. The key assumption is thus that the effect of class size works through educational attainment. Their procedure amounts

2. Chetty et al. (2011) do not only use the STAR experiment to examine the long-term effects of class size; they also investigate the long-term impact of other characteristics of the class in which people where placed in grades K–3.

to imputing the effect of class size on wages using the estimate of class size on schooling and the correlation between schooling and wages. They find that a reduction in the pupil/teacher ratio by one student improves wages by 0.3%.[3]

Although most previous studies are suggestive of a negative long-term effect of class size on adult earnings, the evidence reported therein is by no means conclusive. First, there is considerable uncertainty about the reliability of the imputation approaches. The imputation approach relies on the assumptions that the association between test scores (or schooling) and earnings has a causal interpretation and that the effect of class size on earnings only works through observed test scores or educational attainment. Second, the identifying assumptions in the literature on school quality on earnings are not completely credible. Third, when a credible identification strategy has been used (see Chetty et al. 2011), earnings are observed too early to provide a reliable estimate of the long-run impact of class size on labor market success.

Using unique Swedish data, we trace the effects of changes in class size in primary school on cognitive and noncognitive achievement at ages 13, 16, and 18, as well as on long-term educational attainment, wages, and earnings observed when individuals are aged 27–42. We exploit variation in class size attributable to a maximum class size rule in Swedish primary schools. This maximum class size rule gives rise to a (fuzzy) regression discontinuity design. We apply this identification strategy to data covering the cohorts born in 1967, 1972, 1977, and 1982. The focus on these cohorts is motivated by the fact that we have information on cognitive and noncognitive achievement at the end of primary school for a 5%–10% sample of these cohorts. To these data we match individual information on educational attainment and earnings. Educational attainment and earnings are observed in 2007–2009.

We find that smaller classes in the last three years of primary school (age 10 to 13) are beneficial for cognitive and noncognitive test scores at age 13 and for achievement test scores at ages 16. We also document improvements in noncognitive tests (which are only available for men) at age 18. Most important, we find that smaller classes increase completed education, wages, and

3. Bingley, Jensen, and Walker (2005) apply a similar imputation method using Danish data.

earnings at age 27 to 42. We compare the direct estimate of the wage effect to estimates obtained using the imputation methods of previous studies, and find that the direct estimate is substantially larger than any imputed estimate. A cost-benefit analysis suggests that a reduction in class size from 25 to 20 pupils has an internal rate of return of almost 18%.

The paper proceeds as follows. In Section II we describe the relevant institutions of the Swedish schooling system. Section III describes the data and Section IV the estimation strategy. Section V presents results concerning the validity and strength of our instrumental variable approach, and Section VI presents and discusses the empirical findings. Section VII summarizes and concludes.

II. Institutional Background

In this section we describe the institutional setting pertaining to the cohorts we are studying (the cohorts born 1967–1982). During the relevant time period, earmarked central government grants determined the amount of resources invested in Swedish compulsory schools, and allocation of pupils to schools was basically determined by residence.[4] Compulsory schooling was (and still is) nine years. The compulsory school period was divided into three stages: lower primary school, upper primary school, and lower secondary school. Children were enrolled in lower primary school from age 7 to 10 where they completed grades 1 to 3; after that they transferred to upper primary school where they completed grades 4 to 6. At age 13 students transferred to lower secondary school.

The compulsory school system had several organizational layers. The primary unit in the system was the school. Schools were aggregated to school districts (note that these school districts are very different from U.S. school districts).[5]

4. This changed in the 1990s with the introduction of decentralization and school choice. From 1993 onwards compulsory schools are funded by the municipalities; see Björklund et al. (2005) for a description of the Swedish school system after decentralization. Du Rietz, Lundgren, and Wennås (1987) contains an excellent description of the school system prior to decentralization; we base this section on their description.

5. We use the term "school district" for want of a better word. The literal translation from Swedish would be "principal's district" (*Rektorsområde*). The prime responsibility of the school district was to allocate teachers over classes within

School districts typically had one lower secondary school and at least one primary school. The catchment area of a school district was determined by a maximum traveling distance to the lower secondary school. The recommendations concerning maximum traveling distances were stricter for younger pupils, and therefore there were typically more primary schools than lower secondary schools in the school district. There was at least one school district in a municipality.

The municipalities formally ran the compulsory schools. But central government funding and regulations constrained the municipalities substantially. The municipalities could top up on resources given by the central government, but they could not employ additional teachers. The central government introduced county school boards in 1958 to allocate central funding to the municipalities. In addition, the county school boards inspected local schools.[6]

Maximum class size rules have existed in Sweden in various forms since 1920. Maximum class sizes were lowered in 1962, when the compulsory school law stipulated that the maximum class size was 25 at the lower primary level and 30 at the upper primary and lower secondary levels.[7]

We focus on class size in upper primary school, that is, grades 4 to 6. More precisely, the main independent variable in our analyses is the average of the class sizes students experience in grades 4, 5, and 6.[8] The main reason for this focus is data availability. We have more reliable information on schools (and hence school districts) attended for upper primary school than for lower primary school.

The maximum class size rule at the upper primary level stipulated that classes were formed in multiples of 30; 30 students

district. Unlike U.S. school districts, they cannot raise funding on their own and there is no school board. In the Swedish context, the municipality is the closest analogy to U.S. school districts.

6. In the late 1970s, Sweden was divided into 24 counties and around 280 municipalities.

7. The fine details of the rule were changed in 1978. Prior to 1978, the rule was formulated in terms of maximum class size. From 1978 onward, a resource grant (the so-called base resource) governed the number of teachers per grade level in a school. The discontinuity points were not changed.

8. Hence, if a student is in a class of 25 pupils in grade 4, in a class of 24 students in grade 5 and in a class of 23 students in grade 6, the average class size to which this student was exposed in second stage primary school equals 24 ($= \frac{(25+24+23)}{3}$).

in a grade level in a school yielded one class, while 31 students in a grade level in a school yielded two classes, and so on.[9] We use this rule for identification in a (fuzzy) regression discontinuity (RD) design. This method has been applied in several previous studies to estimate the causal effect of class size.[10]

Implementing the RD design must be done with care, however. The compulsory school law from 1962 opened up for adjustment of school catchment areas within school district such that empty class rooms would be filled. In that process, the county school boards were instructed to take the "needs" of the pupil population into account. Thus, it is likely that the school catchment areas are adjusted within school districts to favor disadvantaged pupils. In a companion paper we show that such sorting takes place, rendering the RD design at the school level invalid.[11] Because of these problems, we implement the RD design at the school district level. The virtue of the school district level is that pupils were assigned to a school district given their residential address and district boundaries were not adjusted in response to enrollment levels. A problem with the school district analysis is that the maximum class size rule has less bite in districts with more than one school. For that reason we focus on districts containing one upper primary school, which we refer to as one-school districts. We provide evidence that the RD design at the school district level is valid in Section V.

The RD design requires, inter alia, that other school resources do not exhibit the same discontinuous pattern. There is no such pattern. In the mid-1980s, for instance, central government money for teachers amounted to 62% of the overall grant. The only other major grant component (27% of the grants) was aimed at supporting disadvantaged students. This grant was tied to the overall

9. There have always been special rules in small schools. In such areas, the rules pertained to total enrollment in two or three grade levels.

10. The seminal paper is Angrist and Lavy (1999). See also Gary-Bobo and Mahjoub (2006); Hoxby (2000); Leuven, Oosterbeek, and Rønning (2008); Urquiola and Verhoogen (2009).

11. In Fredriksson, Öckert, and Oosterbeek (2012) we show that there is bunching around the cutoffs when school enrollment is the forcing variable. In particular it is more likely that schools are found just below than just above the cutoffs. Moreover, expected class size according to the rule predicts parental education; more children with well-educated parents are found just below the kink when school enrollment is the forcing variable.

number of compulsory school students in a municipality and there were no discontinuities in the allocation of the grant.

III. DATA

III.A. Data Sources and Definitions of Key Variables

The key data source is the so-called ETF project which is run by the Department of Education at Göteborg University; see Härnquist (2000) for a description of the data. Among other things, the data contain cognitive test scores at age 13 for roughly a 10% sample of the cohorts born 1967, 1972, and 1982. In addition, there is information on a 5% sample for the cohort born in 1977. For all cohorts, a two-stage sampling procedure was used. In the first stage, around 30 out of the 280 municipalities were systematically selected; the selection criteria were based on, for example, population size and political majority. In the second stage, classes were randomly sampled within municipality. This sampling procedure implies that comparisons across municipalities for a given cohort are not valid, but comparisons within municipalities are valid. For this reason all analyses condition on municipality-by-cohort fixed effects.

To these data we have matched register information maintained by Statistics Sweden. The added data include information on class size (from the Class register), parental information (which is made possible by the multigenerational register containing links between all parents and their biological or adopted children), and medium-term and long-term outcomes. Class size is measured at the school by grade level. The medium-term outcomes are achievement test scores (at age 16) and scores on cognitive and noncognitive tests (at age 18). Long-term outcomes are completed education, earnings and wages measured in 2007-2009. The cognitive and noncognitive test scores at age 18 are only available for men because they are derived from the military enlistment.

The cognitive tests at age 13 are traditional IQ-type tests. We construct a measure of cognitive ability based on scores for verbal skills and logical skills.[12] The verbal test involves finding an

12. We focus on these skills because they are readily comparable to the achievement tests at age 16. There is also information on spatial ability in the data. Including spatial ability in the measure of cognitive ability produces a slightly

antonym of a given word. The logical test requires the respondent to fill in the next number in a sequence of numbers. We standardize cognitive ability such that the mean is 0 and the standard deviation equals 1. The measure of noncognitive skills at age 13 is based on a questionnaire about the pupils' situation in school. We form an index based on questions on, for example, effort, motivation, aspirations, self-confidence, sociability, absenteeism, and anxiety. The construction of the index is described more fully in the Online Appendix. The index is standardized to mean 0 and standard deviation 1.

Academic achievement at age 16 is measured as test scores at the end of lower secondary school. The achievement tests involve maths and Swedish.[13] These achievement tests were used to anchor subject grades at the school level: The school average test result thus determined the average subject grade at the school level. Also this outcome is standardized to mean 0 and standard deviation 1.

The military enlistment cognitive test is very similar in nature to the test administered at age 13; see Mårdberg and Carlstedt (1993) for a description of the Swedish military enlistment battery. It is designed to measure general ability and it is similar to the AFQT (Armed Forces Qualifications Test) used in the United States. We again construct a standardized measure based on the verbal and logical parts of this test. Upon enlistment, army recruits also have a 20-minute interview with a psychologist who assesses their noncognitive functioning. Details of the psychologists' assessments are classified and we have access to an overall score for noncognitive ability. A recruit is given a high score if considered emotionally stable, persistent, socially outgoing, willing to assume responsibility, and able to take initiatives. Motivation for doing the military service is, however, explicitly not a factor to be evaluated.

Data on educational attainment come from the Educational Register maintained by Statistics Sweden. This register records

lower coefficient on class size. With spatial ability included we obtain an estimate of −0.030 (standard error: 0.014) which should be compared to −0.033 (standard error: 0.015).

13. There is also information on an English test. We focus on maths and Swedish since they are readily comparable to the two IQ tests. The estimates are slightly lower if we include English in the measure of achievement at age 16, but still statistically significant.

the highest attained education level for the resident population.[14] We construct two measures based on this. The first is years of completed schooling, which is inferred from the highest level attained.[15] The second is a binary indicator for having at least a bachelor's degree, which is analogous to the college indicator used in studies based on the STAR experiment (see Krueger and Whitmore 2001; Schanzenbach 2007; Chetty et al. 2011). Data on annual earnings come from the Income Tax Register, and data on wages stem from the Wage Register; both registers are maintained by Statistics Sweden. Earnings are based on income statements made by employers. The wage data relate to those who are employed on a day of measurement (in September–November) in a particular year and are measured in full-time equivalent wages.[16] We use earnings and wage data from 2007–2009; individuals of the oldest (1967) cohort are then 42 years old and individuals of the youngest (1982) cohort are 27 years old. Earnings and wages are therefore measured at an age when they correlate highly with lifetime income (Böhlmark and Lindquist 2006).

III.B. Identifying Variation and Some Descriptives

As explained in Section II, we must conduct the analysis at the school district level. To avoid problems associated with (lack of) instrument relevance, we focus on districts with one school—one-school districts. Appendix Table A.1 reports descriptive statistics separately for the one-school districts and the full sample.[17] One-school districts include 27% of the school districts in our original sample (the full sample has 697 school districts and 191 of those are one-school districts), and all types of municipalities are represented in both samples. There are some differences across the samples: One-school districts are, of course, smaller in terms

14. The register is complete for individuals with an education from Sweden. Information for immigrants stems from separate questionnaires to new arrival cohorts. The underlying data include information on the courses taken at the university level, which implies that this is a relatively accurate measure of years of schooling even for those who do not have a complete university degree.

15. For further details see the Online Appendix.

16. The wage data are collected by sampling of small firms in the private sector. For large firms in the private sector, and for the entire public sector, the wage data cover all employed individuals.

17. We drop a few school districts where enrollment in grade 4 was too low to pass the formal requirements for forming a class consisting of only one grade level.

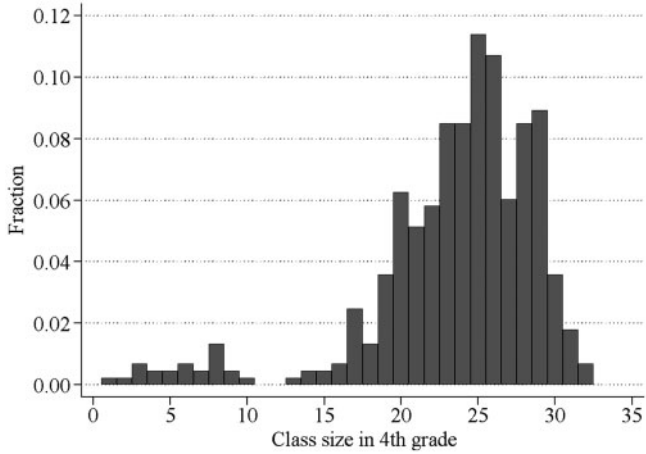

FIGURE I

Distribution of Class Size in Grade 4

The figure shows the distribution of class size in grade 4 in one-school districts for cohorts born 1967, 1972, 1977, and 1982.

of the number of students; and parental education is 0.2–0.3 years higher in the one-school districts than in the full sample. But, overall, the differences are minor: for instance, adult wages differ by 0.8% across samples. We should thus expect the results for the one-school districts to be representative for the results in the full sample.

The second part of Appendix Table A.1 shows that, in the one-school districts, average class size in grades 4–6 is 24.4 pupils. Figure I shows the distribution of class size in grade 4. There are few very small classes (below 15) and few classes (2%) exceed the official maximum class size of 30. The modal class size is 25.

Figure II illustrates the relations between school district enrollment in 4th grade in one-school districts on the horizontal axis, and actual and expected class size in grade 4 on the vertical axis.[18] The solid line shows expected class size, that is, class size in case it would be entirely determined by the maximum class size rule; the dashed line pertains to actual class size. Actual and

18. School enrollment and school district enrollment is obviously the same thing in one-school districts. We use school district enrollment to emphasize that enrollment is not manipulated.

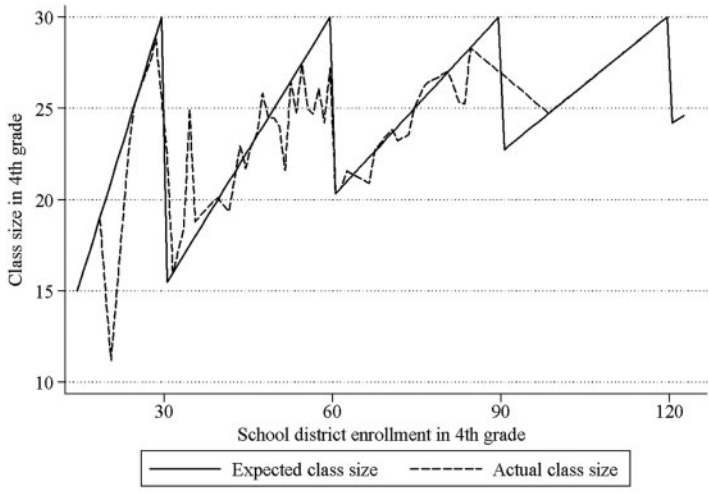

FIGURE II

Expected and Actual Class Size in Grade 4 by Enrollment in Grade 4

The figure shows expected and actual class size in grade 4 by enrollment in grade 4 in one-school districts for cohorts born 1967, 1972, 1977, and 1982. The figure only includes enrollment counts with at least two school districts. After enrollment of 100 students, actual class size exactly coincides with expected class size.

expected class sizes move fairly closely together. (The peak at enrollment just above 30 is caused by one age-integrated class and the two lines exactly coincide for enrollment above 100.)

Table I shows the number of one-school districts by cohort in the vicinity (defined by the window width) of the 1st–4th thresholds. The number of observations in the neighborhood of each threshold is too small to estimate the effect of larger classes at each separate threshold. We therefore pool the data from the different thresholds. In the next section we outline how we approach this.

IV. ESTIMATION STRATEGY

To gain precision we pool the data from the different enrollment thresholds in the following way. Define the thresholds, $\bar{E}_\tau$, as $\bar{E}_\tau = \{30, 60, 90, 120\}$ and the indicator variable $I_{d\tau} = I(E_d \in \bar{E}_\tau \pm W)$. Thus $I_{d\tau} = 1$ if district enrollment (d

TABLE I
NUMBER OF ONE-SCHOOL DISTRICTS BY COHORT AROUND THE ENROLLMENT THRESHOLDS

	Window width		
	±1	±5	±15
1st threshold (enrollment = 30)	3	24	56
2nd threshold (enrollment = 60)	6	31	100
3rd threshold (enrollment = 90)	2	5	28
4th threshold (enrollment = 120)	1	4	7
All thresholds	12	64	191

Notes. The table reports the number of one-school districts by cohort where the enrollment count is at most 1, 5, or 15 away from a threshold for cohorts born 1967, 1972, 1977, and 1982.

indexes school districts) belongs to segment τ, where each segment is defined as enrollment counts within $\pm W$ of $\bar{E}_\tau$. Our default specification has $W = 15$, but conceptually $W = 1, \ldots, 30$.[19] With four different enrollment thresholds and $W = 15$, there are 120 potential enrollment counts. The actual number of enrollment counts with at least one observation is smaller (77). Especially for enrollment counts above 90 the distribution is thin (see Figure III).

Define normalized enrollment as $e_{d\tau} = (E_d - \bar{E}_\tau)I_{d\tau}$ and the treatment indicator

$$Above_{d\tau} = I(e_{d\tau} > 0), \tag{1}$$

For an individual i, the outcome equation of interest is

$$y_{id\tau} = \beta CS_{d\tau} + \alpha_\tau + f^k_\tau(e_{d\tau}) + \in_{id\tau}, \tag{2}$$

where we use $Above_{d\tau}$ as the instrument for class size ($CS_{d\tau}$):

$$CS_{d\tau} = \gamma Above_{d\tau} + \delta_\tau + g^k_\tau(e_{d\tau}) + \nu_{d\tau}, \tag{3}$$

To accommodate different patterns around different thresholds, we include segment fixed effects (α_τ and δ_τ) and allow the coefficients on the enrollment polynomials of degree k ($f^k_\tau(e_{d\tau})$ and $g^k_\tau(e_{d\tau})$) to vary by segment. This approach parallels analyses of

19. With $W > 15$, the same observation is used as treated for one threshold and control for the next. For example, for $W \geq 17$, a district with enrollment equal to 47 belongs to the treated group at the threshold of 30 and to the control group at the threshold of 60.

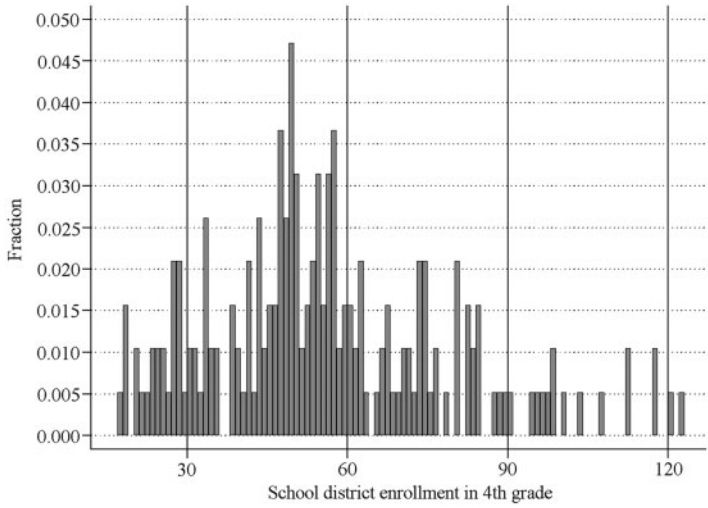

FIGURE III

Distribution of Enrollment in Grade 4 in One-School Districts

The figure shows the distribution of enrollment in grade 4 in one-school districts for cohorts born 1967, 1972, 1977, and 1982.

randomized experiments with conditional random assignment (e.g., Krueger 1999; and Black et al. 2003), where each threshold is regarded as a different experiment.[20]

Notice that the endogenous variable in our analysis is the average of the class sizes student experience in grades 4, 5, and 6, while the instrument is derived from enrollment in grade 4. There are two reasons for this. The first reason is that enrollment in fifth and sixth grade are potentially endogenous to class size in fourth grade. Therefore, we cannot validly treat enrollment in fifth and sixth grade as exogenous. Enrollment in fourth grade can arguably be treated as exogenous because third (lower primary school) and fourth grade (upper primary school) belong to different stages of compulsory school. The transition between lower primary and upper primary school often implies a change of school, and class size rules are different in lower primary and

20. Potentially, there would be an efficiency gain of using information on treatment intensity, which varies since the sizes of the jumps in expected class size vary across segments. Column (2) of Table AV in the Online Appendix effectively explores such a specification. In practice, the efficiency gain appears limited.

upper primary school. Given that enrollment in fifth and sixth grade are potentially endogenous, we have no instruments for class size in grades 5 and 6. The second reason is that class sizes in grades 4, 5, and 6 are highly correlated. The correlation between class size in grades 4 and 5 is .79 and the correlation between class size in grades 4 and 6 is .57. Attributing all effects only to class size in grade 4 would not be correct. By focusing on the average of the class sizes in grades 4, 5, and 6, the instrumental variables (IV) estimates reflect the effects of an increase of class size by one pupil during three years.

V. Validity of the Instrument

A threat to the validity of the RD design is bunching on one side of the cutoffs, since that indicates that the forcing variable is manipulated. Urquiola and Verhoogen (2009) document an extreme example of bunching in the context of a maximum class size rule in Chile. In their data there are at least five times as many schools just below than just above the cutoffs.

Figure III shows the distribution of enrollment in grade 4 in one-school districts. Visual inspection reveals no suspect discontinuities in the distribution of the forcing variable. The McCrary (2008) density test confirms this: We cannot reject the hypothesis that there is no shift in the discontinuity.[21]

A more direct way to assess whether the instrument is valid is to examine if predetermined characteristics are balanced across observations above and below the thresholds. Figure IV shows that this is the case for parental education: The estimated discontinuity is −0.076 with a standard error of 0.369. Analogous plots for other covariates show very similar pictures.

Table II addresses the question of the balancing of predetermined covariates more formally. The first two columns show that the baseline covariates we consider are highly relevant predictors of cognitive ability at age 13 and adult wages (observed at age 27–42). For instance, children who have more educated mothers score higher on the cognitive test (a year of education is associated with an increase in test scores of 0.069 standard

21. To implement the test we used a bin size of one student and a bandwidth of five students. The estimated log difference in the height of the density is 0.19 with a standard error of 0.57.

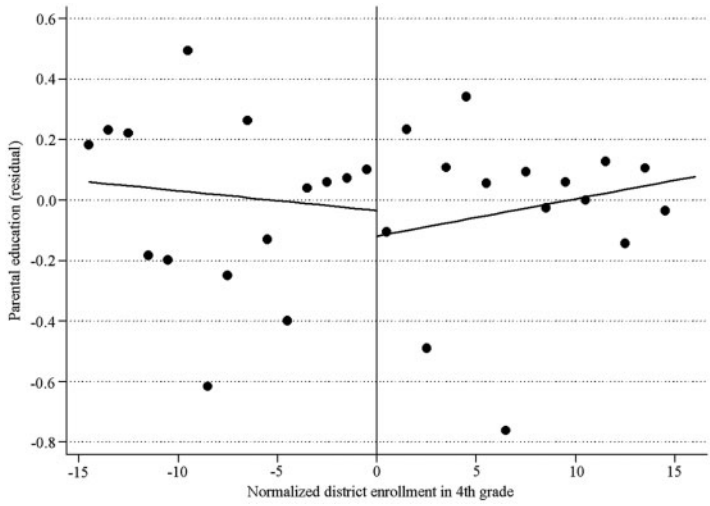

FIGURE IV

Parental Education by Enrollment in Grade 4

The figure shows residual parental education, after controlling for fixed effects for enrollment segments and municipality-by-cohort fixed effects, by normalized enrollment in grade 4. The data pertain to one-school districts for cohorts born 1967, 1972, 1977, and 1982. The regression lines were fitted to individual data. Discontinuity at threshold: −0.076 (standard error: 0.369).

deviations) and go on to have higher wages (a year of education is associated with a 0.6% increase in wages).

Column (3) of Table II shows the result of regressing the instrument on all baseline covariates.[22] The next to last row contains the result of an F-test of the hypothesis that all the coefficients on baseline covariates are jointly zero. The message of this F-test is that pre-determined characteristics are unrelated to the instrument (the p-value is .70). In column (4) we test whether the coefficient on the instrument is different from zero in a regression of each individual characteristic on the instrument. Again, predetermined characteristics are unrelated to the instrument.

22. These results come from regressions where we control for enrollment segment fixed effects; linear controls for normalized school district enrollment, where the slopes are allowed to differ above and below the thresholds as well as across segments; and municipality-by-cohort fixed effects. We justify this specification in detail shortly.

TABLE II
BALANCING OF COVARIATES

	(1) Cognitive ability age 13	(2) ln(Wage) age 27–42	(3) Above threshold	(4) *p*-value
Female	−0.0020	−0.1422***	0.0027	.433
	(0.0253)	(0.0107)	(0.0035)	
Month of birth	−0.0227***	−0.0017	0.0005	.453
	(0.0035)	(0.0014)	(0.0007)	
Immigrant	−0.4616***	0.0222	0.0113	.566
	(0.0585)	(0.0226)	(0.0168)	
Mother's years of education	0.0687***	0.0064**	0.0004	.981
	(0.0059)	(0.0023)	(0.0010)	
Father's years of education	0.0597***	0.0135***	−0.0009	.665
	(0.0051)	(0.0018)	(0.0010)	
Parental income (SEK 100,000s)	0.0384***	0.0112***	0.0002	.947
	(0.0074)	(0.0026)	(0.0018)	
Mother's age at birth	0.0189***	0.0027***	−0.0004	.471
	(0.0023)	(0.0009)	(0.0007)	
Number of siblings	−0.0728***	−0.0057*	−0.0011	.709
	(0.0116)	(0.0045)	(0.0022)	
Parents separated	−0.1066***	−0.0305**	−0.0053	.580
	(0.0299)	(0.0118)	(0.0057)	
p-value of *F*-test	.000	.000	.698	
Number of individuals	5,116	3,185	5,920	

Notes. The estimates are based on representative samples of individuals born in 1967, 1972, 1977 or 1982 in one-school districts. Cognitive ability at age 13 is standardized. The ln(wage) estimates are restricted to wage-earners. Columns (1)–(3) report results of OLS regressions on the variables listed in the rows. These regressions also include the following control variables: fixed effects for enrollment segment, linear controls for school district enrollment interacted with threshold and segment, and municipality-by-cohort fixed effects. Above threshold (the instrument for class size) is an indicator equalling unity if school district enrollment in fourth grade exceeds the class size rule threshold in the enrollment segment. Independent variables are predetermined parent and student characteristics. The *p*-value reported at the bottom of columns (1)–(3) is for an *F*-test of the joint significance of the variables listed in the table. Each row of column (4) reports a *p*-value from separate OLS regressions of the predetermined variable (listed in the corresponding row) on the instrument, and the same set of control variables as in columns (1)–(3). The *p*-value is for a *t*-test of the significance of the class size instrument. Standard errors adjusted for clustering by enrollment count (77 clusters) are in parentheses Asterisks indicate that the estimates are significantly different from zero at the ***1% level, **5% level, and *10% level.

Figure V shows a graphical representation of the first stage. There is a clear, and statistically significant, jump at the threshold. Districts that have just surpassed one of the thresholds have classes that are systematically smaller than classes just below the threshold. The discontinuity at the threshold is −5.21 with a standard error of 0.85. When we control for predetermined characteristics, which is valid given the results in Table II, the

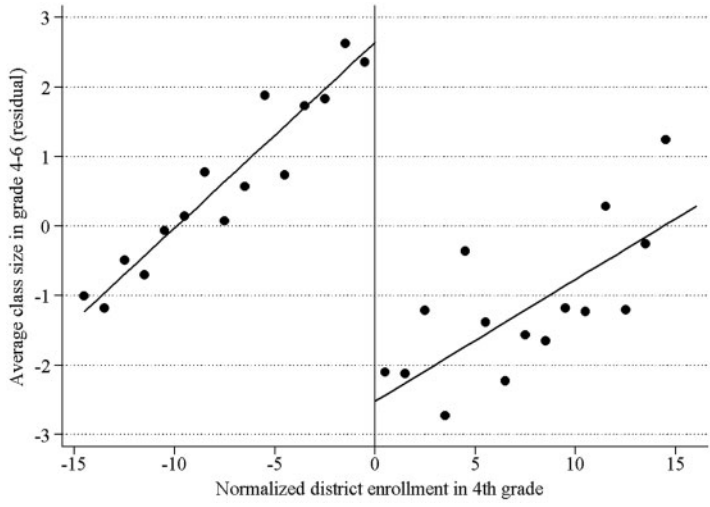

FIGURE V

Class Size by Enrollment in Grade 4

The figure shows residual average class size in grades 4–6, after controlling for fixed effects for enrollment segments and municipality-by-cohort fixed effects, by normalized enrollment in grade 4. The data pertain to one-school districts for cohorts born 1967, 1972, 1977, and 1982. The regression lines were fitted to individual data. Discontinuity at threshold: −5.207 (standard error: 0.848).

estimate of the discontinuity at the threshold does not change at all (the estimate is −5.22, with a standard error 0.85).

VI. THE EFFECTS OF CLASS SIZE

We start with a graphical analysis for a subset of our outcome variables. To improve precision, we examine the residuals from regressions where we control for predetermined characteristics (and municipality-by-cohort fixed effects). Figure VI shows average cognitive ability at age 13 by one-student bins. There is a clear discontinuity at the threshold. School districts having surpassed one of the thresholds (that on average have smaller classes) score better on the cognitive tests at age 13. Figure VII presents the analogous picture for wages at age 27–42. Again,

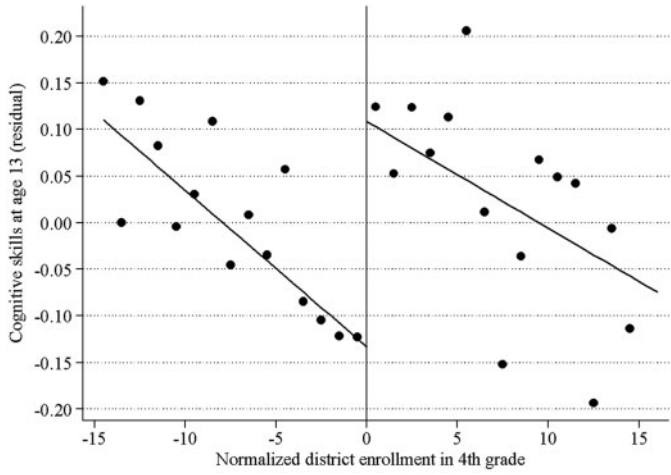

FIGURE VI

Cognitive Ability at Age 13 by Enrollment in Grade 4

The figure shows residual cognitive ability, by normalized enrollment in grade 4. The residual comes from a regression of cognitive ability on fixed effects for enrollment segments, municipality-by-cohort fixed effects, gender, dummy variables for month of birth, dummy variables for mother's and father's educational attainment, parental income, mother's age at child's birth, indicators for being a first- or second-generation Nordic immigrant, indicators for being a first- or second-generation non-Nordic immigrant, an indicator for having separated parents, and the number of siblings. The data pertain to one-school districts for cohorts born 1967, 1972, 1977 and 1982. The regression lines were fitted to individual data. Discontinuity at threshold: 0.244 (standard error: 0.076); without baseline covariates, the estimate is 0.252 (standard error: 0.135).

outcomes improve (i.e. wages increase) in districts that have just surpassed the thresholds.[23]

The remainder of this section quantifies the jumps at the thresholds using regression analysis. The regressions have the same basic structure as equations (2) and (3). In the regressions we include baseline covariates (linearly) to improve precision and municipality-by-cohort fixed effects.[24] Throughout, the

23. Figure AII in the Online Appendix shows similar figures where we do not control for predetermined characteristics.

24. For a given cohort, the effects are thus identified from municipalities with at least two one-school districts. If we exclude the 24 school districts that do not contribute to identification, the estimates do not change at all.

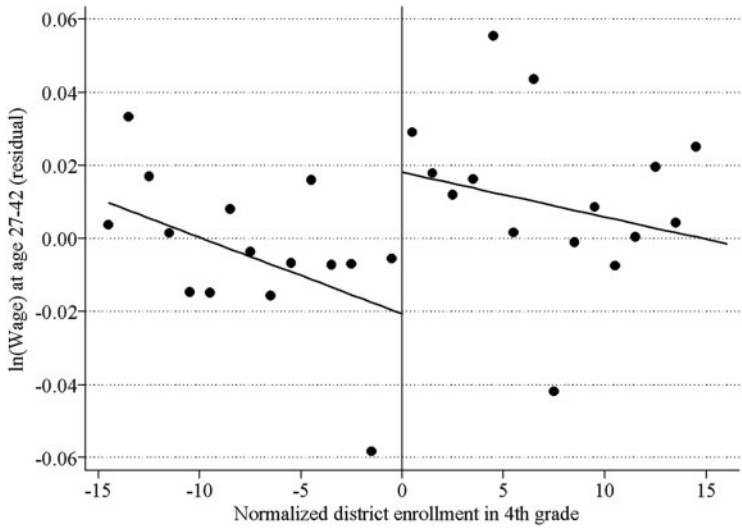

FIGURE VII

Adult Wages by Enrollment in Grade 4

The figure shows residual wages, by normalized enrollment in grade 4. The residual comes from a regression of log wages on fixed effects for enrollment segments, municipality-by-cohort fixed effects, gender, dummy variables for month of birth, dummy variables for mother's and father's educational attainment, parental income, mother's age at child's birth, indicators for being a first- or second-generation Nordic immigrant, indicators for being a first- or second-generation non-Nordic immigrant, an indicator for having separated parents, and the number of siblings. The data pertain to one-school districts for cohorts born 1967, 1972, 1977, and 1982. The regression lines were fitted to individual data. Discontinuity at threshold: 0.039 (standard error: 0.016); without baseline covariates, the estimate is 0.029 (standard error: 0.020).

regression error terms are clustered by school district enrollment count as suggested by Lee and Card (2008).[25]

VI.A. Specification Analysis

Table III shows IV estimates of the effect of class size on cognitive skills and wages. In addition we provide information on the corresponding reduced-form (RF) and first-stage estimates. The RF equations regress the outcomes on the binary

25. Clustering on the enrollment counts yields 77 clusters. This is a higher level than the school district by cohort level (191 clusters) which is the level where our instrument varies. Table AIII in the Online Appendix shows that using different levels of clustering has only minor effects on the standard errors.

TABLE III

REDUCED-FORM (RF) AND IV ESTIMATES, DIFFERENT ENROLLMENT CONTROLS

Model	(1)	(2)	(3)	(4)	(5)	(6)
	Cognitive ability, age 13 ($N = 5{,}116$)					
RF: Above threshold	0.2546***	0.2405***	0.2493***	0.2089**	0.2144**	0.3858***
	(0.0732)	(0.0771)	(0.0766)	(0.0908)	(0.0876)	(0.1223)
IV: Class size grades 4–6	−0.0471***	−0.0457***	−0.0454***	−0.0317*	−0.0330**	−0.0628**
	(0.0163)	(0.0165)	(0.0161)	(0.0147)	(0.0146)	(0.0292)
	ln(Wage), age 27–42 ($N = 3{,}185$)					
RF: Above threshold	0.0415***	0.0362**	0.0425***	0.0343	0.0428**	0.0761***
	(0.0148)	(0.0178)	(0.0149)	(0.0230)	(0.0212)	(0.0282)
IV: Class size grades 4–6	−0.0070***	−0.0062*	−0.0071***	−0.0048	−0.0063*	−0.0114**
	(0.0026)	(0.0032)	(0.0026)	(0.0033)	(0.0032)	(0.0057)
	Average class size grades 4–6 (first stage) ($N = 5{,}920$)					
Above threshold	−5.4215***	−5.2766***	−5.5303***	−6.7143***	−6.6254***	−6.3740***
	(0.8899)	(0.8715)	(0.8729)	(0.8064)	(0.7523)	(1.4522)
F-test for instrument	37.12	36.65	40.14	69.32	77.56	19.26
Enrollment controls						
1st-order polynomial	✓		✓		✓	
2nd-order polynomial		✓		✓		✓
Interacted with segments			✓	✓	✓	✓
Interacted with thresholds					✓	✓
Number of districts × cohorts	191	191	191	191	191	191

Notes. The estimates are based on representative samples of individuals born in 1967, 1972, 1977 or 1982 in one-school districts. Cognitive ability at age 13 is standardized. The wage measure is an average during 2007–2009 and is restricted to wage-earners. Above threshold (the instrument for class size) is an indicator equaling unity if school district enrollment in fourth grade exceeds the class size rule threshold in the enrollment segment. RF refers to reduced form, where outcomes are regressed on the indicator for being above the class size rule threshold. IV refers to instrumental variables, where average class size in grades 4–6 is instrumented by the indicator for being above the class size rule threshold. In addition to the control variables listed in the table, all models include fixed effects for enrollment segments, municipality-by-cohort fixed effects, gender, dummy variables for month of birth, dummy variables for mother's and father's educational attainment, parental income, mother's age at child's birth, indicators for being a first- or second-generation Nordic immigrant, indicators for being a first- or second-generation non-Nordic immigrant, an indicator for having separated parents and the number of siblings. Standard errors adjusted for clustering by enrollment count (77 clusters) are in parentheses. Asterisks indicate that the estimates are significantly different from zero at the ***1% level, **5% level, and *10% level.

indicator for being above a threshold, the enrollment polynomial, and other controls variables. We provide these estimates for six different specifications of the enrollment polynomials $f_\tau^k(\cdot)$ and $g_\tau^k(\cdot)$. Columns (1) and (2) restrict the polynomials to be the same across segments, columns (3) and (4) allow the polynomials to differ across segments, and columns (5) and (6) interact the polynomials with segment and threshold. Conceptually, we favor the specifications in columns (3)–(6). These specifications account for the differences in slopes of the expected class size function across segments. We also have a slight preference for the more flexible specification in columns (5) and (6) over the specification in columns (3) and (4).

The first five columns suggest that the results are stable. The RF effect on cognitive ability varies between 0.21 and 0.25 of a standard deviation and the corresponding IV estimates suggest an impact ranging between −0.032 and −0.047 standard deviation units per student increase in class size. The reduced form effect on wages varies between 3.4% and 4.3% percent and the IV estimates on wages correspond to a reduction of 0.5% to 0.7% percent per unit increase in class size.

It is useful to compare these IV estimates of class size to the corresponding ordinary least squares (OLS) estimates. The OLS estimate of class size on cognitive ability is a precisely determined zero: the estimate is 0.003 (with a standard error of 0.007). The OLS estimate of class size on log wages is 0.002 (with a standard error of 0.002). The OLS estimates are obviously biased upward, suggesting that there is compensatory allocation of class size.

The results in column (6), which has a quadratic in enrollment interacted with segment and threshold, are very different from the results in the previous five columns. There are sharp (and to our minds unrealistic) increases in the RF effects (they are still statistically significant, however). We also observe a substantial drop in the power of the instrument: when we move from column (5) to column (6), the *F*-statistic drops from 78 to 19. It seems that the specification in column (6) is too flexible relative to the identifying variation in the data.

Based on the evidence in Table III, we think that the specification with a linear enrollment control which is interacted with threshold and segment is the most sensible specification. Notice that we have also tested for the optimal order of the polynomial using the Akaike information criteria as suggested by Lee and Lemieux (2010). It turns out that the optimal polynomial order

for the outcomes implies a less flexible model. This result is probably driven by the fact that we have a relatively limited number of school districts in our data. We prefer to be consistent with what we see in the graphs rather than relying on the Akaike test. According to the graphs, the linear interacted polynomial seems like the most sensible specification.

A related issue is bandwidth selection. The regression results in Table III are based on bandwidths of ± 15 students around the thresholds. What if we reduce the bandwidth? Table AIV in the Online Appendix shows reduced form estimates when the bandwidth is increased by two students from ± 7 to ± 15. We start at ± 7 to have a decent amount of districts (90) and clusters (38) to identify the effects from.[26] The table shows that the RF effects on cognitive ability and wages are statistically significant even at the smaller bandwidths, and that the estimates for the smaller bandwidths are not statistically different from the baseline estimates, which are based on ± 15 students around the thresholds.

VI.B. The Exclusion Restriction

An important question is whether the instrument only affects the outcomes via its effect on class size. If this is not the case, we can only causally interpret the RF estimates shown in Table III. To provide evidence on the validity of the exclusion restriction, we examine if districts respond in other ways to the class size rule. Results are presented in columns (1)–(6) of Table IV. In column (1), we examine whether the probability of being assigned to remedial training is affected by the instrument. If schools respond to the instrument, we would expect it to be lower in districts that have surpassed one of the thresholds. We find no such evidence, however. Column (2) examines if the probability of being assigned to an age-integrated class is affected by the instrument. Again, we find no evidence that this is an issue.

In columns (3) and (4) we examine the possibility that there may be greater scope for tracking when a threshold is surpassed, since surpassing a threshold implies the addition of another class. To address this issue we construct two dissimilarity indices (Duncan and Duncan 1955) which relate class composition to

26. When we reduce the bandwidth to ± 5, there are 64 school districts and 27 clusters. In general, the estimates for the smaller bandwidths should be interpreted with some care since we are asking a lot from the data.

TABLE IV

OTHER RESPONSES TO THE INSTRUMENT

			Class composition		Teacher characteristics		Class size		
	Remedial training	Age integrated class	Education	Income	Experience	Education			
Grade	4	4	4	4	4	4	1–3	4–6	7–9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Above threshold	−0.062	0.054	0.033	0.009	−0.312	−0.036*	−0.745	−6.625***	−0.693
	(0.054)	(0.047)	(0.026)	(0.022)	(0.622)	(0.019)	(1.034)	(0.752)	(0.867)
Number of districts × cohorts	191	191	191	191	191	191	191	191	191
Number of individuals	4,346	5,920	5,920	5,920	5,834	5,834	5,896	5,920	5,920

Notes. The estimates are based on representative samples of individuals born in 1967, 1972, 1977, or 1982 in one-school districts. Remedial training equals one if the pupil attends remedial training, age integrated class is the share of pupils in the school who are placed in an age-integrated class, class composition with respect to education is the dissimilarity index for parental education, class composition with respect to income is the dissimilarity index for parental income, teacher experience is the average years of experience for teachers in grades 4–6 in the school, teacher education is the share of teachers with a college degree in grades 4–6 in the school and class size is the average class size in the grades. Above threshold (the instrument for class size) is an indicator equalling unity if school district enrollment in fourth grade exceeds the class size rule threshold in the enrollment segment. All models include the following controls for school district enrollment in grade 4: fixed effects for enrollment segment; linear controls for enrollment which are interacted with threshold and segment. In addition all models include the following baseline controls: municipality-by-cohort fixed effects, gender, dummy variables for month of birth, dummy variables for mother's and father's educational attainment, parental income, mother's age at child's birth, indicators for being a first- or second-generation Nordic immigrant, indicators for being a first- or second-generation non-Nordic immigrant, an indicator for having separated parents, and the number of siblings. Standard errors adjusted for clustering by enrollment count (77 clusters) are in parentheses. Asterisks indicate that the estimates are significantly different from zero at the ***1% level, **5% level, and *10% level.

school composition. Column (3) considers segregation in terms of parental education and column (4) considers parental income. In both cases segregation is unrelated to the instrument.[27]

In columns (5) and (6) we relate teacher characteristics to the instrument. The rule is unrelated to teacher experience; see column (5). But there is some evidence that the share of teachers with a college degree is lower in districts having surpassed one of the thresholds; see column (6). This may be a source of concern. Note, however, that the reduction in teacher credentials is arguably driven by the decrease in class size; moreover, there is little credible evidence suggesting that teacher credentials affect student performance. Nevertheless, smaller classes come with less educated teachers. If anything, this would tend to reduce our estimate of class size relative to an ideal experiment conducted in our context.

An issue that affects the interpretation of the IV estimates is whether class size in grades 4–6 is correlated with class sizes in the other stages of compulsory school. Columns (7) and (9) in Table IV address this issue by showing results from regressions of class size in lower primary school (grades 1–3) and class size in lower secondary school (grades 7–9) on the instrument. The estimates show that class sizes in the other stages of compulsory school are unrelated to the instrument. Dividing the estimates in column (9) with the first-stage estimate in column (8), we find that a pupil increase in class size in upper primary school leads to an (insignificant) 0.10 increase of class size in lower secondary school. The correlation with class size in lower primary school (obtained analogously) is 0.11, which is also insignificant.

Given the evidence in Table IV, we focus on IV estimates from here on. We interpret these IV estimates as the effects of one pupil change throughout upper primary school (grades 4–6).

VI.C. The Main Results

Table V presents IV estimates of the impact of class size on educational and labor market outcomes observed from age 13

27. Notice that the standard errors are biased downwards in columns (3) and (4). The bias comes from the fact that the indices has complete evenness as the baseline. Since classes are small units, the appropriate baseline is random unevenness. To generate the appropriate baseline one should simulate the baseline by randomly allocating individuals to units; see Carrington and Troske (1997) on these points. Since our estimates are not significant even with complete evenness as the baseline, we have refrained from simulating the data.

TABLE V

IV ESTIMATES OF CLASS SIZE IN FOURTH–SIXTH GRADE

Outcome [# individuals]	One-school districts (1)	One-school districts (2)
Ability measures		
Cognitive ability, age 13	−0.0330**	−0.0327
[$N = 5{,}116$]	(0.0146)	(0.0230)
Noncognitive ability, age 13	−0.0265**	−0.0263**
[$N = 4{,}681$]	(0.0118)	(0.0119)
Academic achievement, age 16	−0.0233**	−0.0211
[$N = 5{,}318$]	(0.0101)	(0.0180)
Educational attainment (age 27–42)		
Years of schooling	−0.0545**	−0.0480
[$N = 5{,}588$]	(0.0256)	(0.0459)
P(bachelor's degree)	−0.0076*	−0.0063
[$N = 5{,}920$]	(0.0043)	(0.0066)
Labor market outcomes (age 27–42)		
Earnings (effect relative to the average)	−0.0117*	−0.0099
[$N = 5{,}920$]	(0.0061)	(0.0066)
ln(wage)	−0.0063*	−0.0043
[$N = 3{,}185$]	(0.0033)	(0.0037)
P(earnings > 0)	−0.0016	−0.0011
[$N = 5{,}920$]	(0.0024)	(0.0029)
Baseline covariates	Yes	No
Number of districts × cohorts	191	191

Notes. The estimates are based on representative samples of individuals born in 1967, 1972, 1977 or 1982 in one-school districts. All ability measures are standardized. The educational outcomes are measured in 2009, while the labor market outcomes have been averaged over the 2007–2009 period. Earnings effects (and their standard errors) are divided by average earnings level to facilitate interpretation. The ln(wage) estimates are restricted to wage-earners. Average class size in grades 4–6 is instrumented with Above threshold (=1 if school district enrollment in fourth grade exceeds the class size rule threshold in the enrollment segment). All models include the following controls for school district enrollment in grade 4: fixed effects for enrollment segment; linear controls for enrollment which are interacted with threshold and segment. In addition all models include the following baseline controls: municipality-by-cohort fixed effects, gender, dummy variables for month of birth, dummy variables for mother's and father's educational attainment, parental income, mother's age at child's birth, indicators for being a first- or second-generation Nordic immigrant, indicators for being a first- or second-generation non-Nordic immigrant, an indicator for having separated parents, and the number of siblings. Standard errors adjusted for clustering by enrollment count (77 clusters) are in parentheses. Asterisks indicate that the estimates are significantly different from zero at the ***1% level, 5% level, and *10% level.

(when the intervention ended) until prime age (age 27–42). Each row refers to a different outcome. Column (1) reports the main results, where we condition on baseline covariates, and column (2) omits the baseline covariates. Throughout we use the specification in column (5) of Table III.

Overall, the point estimates are reassuringly similar across columns. The only real difference between the two columns is that the estimates in column (2) are less precise. Another overall feature is that the estimates in column (1) are remarkably consistent for various outcomes. The effects are always negatively signed and (almost) always statistically significant. The effects on the short and medium term outcomes are significant at the 5% level. The long-term effects are somewhat less precise, but typically significant at the 5% or 10% level.

The first two rows relate to short-term outcomes: cognitive and noncognitive ability measured at the end of primary school when students are 13 years old. The estimate on cognitive ability is negative and significantly different from zero. Placement in a small class during grades 4 to 6 increases cognitive ability at age 13. The estimate suggests that a class size reduction equivalent to STAR (seven students), would improve cognitive skills at age 13 by 0.23 of a standard deviation (SD). The short-run effect on test scores is thus on par with the typical estimate from STAR. Krueger's (1999) estimates of the initial and cumulative effects (his Table IX) translate into an achievement gain of 0.22 of a standard deviation in three years. Moreover, our estimate is of the same magnitude as the estimates reported by Angrist and Lavy (1999) for Israel; it is also comparable to Lindahl (2005), who used a difference-in-differences design to estimate the effect of class size for 6th-graders in Stockholm.

The estimates for noncognitive ability suggests that placement in a small class improves this outcome. The magnitude of the effect is slightly smaller than the effect on cognitive ability. A unit reduction in class size improves noncognitive outcomes by 0.026 of a standard deviation. In the previous literature, there is not much evidence on the relationship between noncognitive outcomes and class size. One exception is Dee and West (2008) who, in their analysis of STAR, find a short-run effect on behavior in the fourth grade but no evidence of an impact on eighth-grade behavior.

The second time we have an outcome measure is at the end of lower secondary school when pupils are 16 years old. This is three years after pupils left primary school. Only academic achievement has been measured, and the results are reported in the third row. The estimated effect is consistently negative. Our baseline estimate in column (1) suggests that reducing class size by one pupil increases academic achievement by 0.023 standard

deviation units. The magnitude of the effect in is only slightly smaller than at age 13. There is thus no evidence of substantive fade-out. Unlike STAR, we find that 70% of the initial effect remains three years after the intervention ended.

The remaining rows in Table V pertain to adult outcomes observed when individuals are aged 27–42. A reduction in class size has a positive effect on educational attainment. The estimate in column (1) indicates that a reduction of class size by one pupil during the last three years of primary school increases years of schooling by 0.05 year (or two-thirds of a month). The effects are also visible at the higher end of the education distribution. A reduction of class size by one pupil increases the probability of having a college degree by 0.8 percentage point. This estimate is in fact larger than the estimates from STAR (e.g., Chetty et al. 2011; Dynarski, Hyman, and Schanzenbach 2011) when evaluated at a reduction by seven students.

The three last rows report estimates of class size on labor market outcomes. The first row shows estimates for annual earnings. We include those with zero earnings, and to facilitate interpretation we divide the estimated effects (and their standard errors) by average earnings in the data. When class size is reduced by one, earnings increase by 1.2% relative to the average. Next we examine log wages (in full-time equivalents). We find a 0.6% increase in wages for a pupil reduction in class size. The final row shows that class size variations have no effect on the probability of working (having positive annual earnings). Since the probability of working is unaffected by variations in class size, the wage effects are not driven by the fact that wages are observed for the selected subsample of workers.[28]

Taken together the effects on the three labor market outcomes also imply that annual hours would increase in response

28. The Online Appendix reports results from various robustness checks. Table AII reports results obtained from the full sample also including districts with more than one school. The point estimates are very close to those reported in Table V but the standard errors are larger. Table AIII reports standard errors obtained from different levels of clustering. This has almost no impact. Table AV shows results obtained using the maximum class size rule instead of the dummy as instrumental variable. Results are very similar. Table AVI reports the results from specifications that omit the municipality-by-cohort fixed effects. This mainly affects precision but not the effect estimates. Figure AI and Table AVII show results for the four thresholds (30, 60, 90, and 120) separately. The patterns are very similar for the first three thresholds; at the fourth (120) there is not sufficient data to achieve identification.

to a reduction in class size. This follows from the fact that the earnings effects are larger in absolute value than the wage effects and that there is no effect on the probability of working.

The findings of significant wage and earnings effects are the most important findings of this article. No previous study has been able to demonstrate significantly negative effects of class size in primary school on adult wages or earnings using a credible identification approach. Chetty et al. (2011) make an attempt to estimate the direct effect of class size on earnings and find no effect. However, they observe earnings at age 27, which is very early on in the labor market career. To test the conjecture that this fact contributes to their finding, we have estimated the earnings effect at age 27 using our data. We find an insignificant effect of −0.4% which is substantially lower than the −1.17% reported in Table V.

VI.D. Implications

1. Comparison with Imputed Estimates. Krueger (2003), Schanzenbach (2007), and Chetty et al. (2011) impute the effect of class size on wage earnings by multiplying the effect of class size on cognitive ability with the cross-sectional correlation between cognitive ability and wage earnings. The purpose of this subsection is to illustrate what we would have concluded had we followed this approach.

To implement this approach, we need estimates of the correlation between cognitive test scores and long-term wage outcomes. Table AVIII in the Online Appendix reports the results of regressions of log wages on cognitive and noncognitive test scores measured at age 13. The correlations between the short-term and the long-term outcomes are high. A standard deviation increase in cognitive test scores is associated with a wage increase of 8.2%. Moreover, if cognitive and noncognitive test scores are included jointly, both are highly significant. A standard deviation increase in cognitive test scores is associated with a wage increase of 7.2%, while a standard deviation increase in the noncognitive test score implies a wage increase of 3.3%.

With these estimates in hand we can implement the two-step approach using our data. We find an imputed wage impact of $-0.033 \times 8.2 = -0.27$ %. When we add the "imputed" impact of noncognitive skills, the estimate increases to $(-0.033 \times 7.2) + (-0.026 \times 3.3) = -0.35$ %. If we instead follow Dustmann, Rajah, and van Soest (2003) and use the impact of

class size on completed years of education in the two-stage procedure, the estimate is −0.22%.[29] All these indirect estimates are substantially (but not significantly) below the estimate of −0.63% per pupil that we find when we estimate the wage effect directly.[30]

Alternatively, we can impute an earnings impact based on the estimates in Table V and in Table AVIII in the Online Appendix. The imputed impact, taking both cognitive and non-cognitive ability into account, is −0.48% per pupil. Again this is substantially lower than the estimate of −1.17% per pupil reported in Table V. Since observed abilities measure limited dimensions of the skills that are priced on the labor market, we think it is natural that all these imputation procedures yield a lower estimate than the direct wage or earnings impacts. Imputed wage or earnings effects thus yield conservative estimates of the long-run effects of class size reductions.

2. Cost-Benefit Analysis. The ultimate question is whether the benefits of a class size reduction outweigh the costs of such an intervention. Important here is that the costs are incurred when children are 10 to 13 years old, while the benefits in terms of wages or earnings only start to accrue when these children are adults and enter the labor market. A cost-benefit analysis shows that for all reasonable discount rates, the present value of the benefits exceeds the present value of the costs. In calculating the benefits we focus on the wage effect, which we treat as permanent in line with previous research. The wage effect is arguably a better estimate of how individuals' productivity is affected by a class size reduction than the earnings effect. The variation in annual earnings reflect preferences and labor supply choices to a greater extent than wages.

Assume average class size during upper primary school is reduced from 25 to 20. This increases the number of teachers from 4 per 100 pupils to 5 per 100 pupils, thereby increasing the per pupil wage costs by 1% of teachers' average wage

29. This combines a class size effect of −0.0545 and a rate of return to education of 4%.

30. A test of the hypothesis that the direct effect (−0.63) equals the imputed effect using cognitive skills only (−0.27) yields a p-value of .16. To test this hypothesis we take the covariances between the various components into account.

during three years.[31] There are also costs involved with overhead and extra classrooms; say that this adds one-third to the extra costs of teachers. The present value of the costs — starting when pupils are 10 years old — is then $\Sigma_{t=0}^{2} 0.01w \frac{(1+\frac{1}{3})}{(1+r)^t}$, where w is the annual wage of a teacher and r the discount rate. Assume further that average wages in the country are approximately equal to the average teacher wage, and that people work from age 21 until age 65.[32] The present value of the benefits is then $\Sigma_{t=10}^{54} \frac{0.0315w}{(1+r)^t}$, where 0.0315 is five times our estimate of the effect of a one pupil reduction of class size on wages. The internal rate of return (the discount rate that equalizes the present values of costs and benefits) is equal to 0.178. For discount rates below this value, the net present value of a five-pupil reduction in class size is positive.

These calculations assume that the same quality teachers can be hired at a constant wage rate, and that the supply of more skilled labor does not affect the wage return to the class size reduction. The internal rate of return would be lower if one of these assumptions does not hold. But even if we double the costs and cut the benefits in half, the internal rate of return is quite high: 0.089. This all implies that in the context of Sweden of the 1980s, a class size reduction in upper primary school would have been a beneficial intervention. Had we based this calculation on the earnings effect, our conclusion would of course be reinforced since the effect on earnings of class size reduction (1.17%) is larger than the wage effect (0.63%).

VI.E. Heterogeneity

To examine whether the effects of class size are heterogeneous, we present results where we have interacted class size with parental income and gender respectively. More precisely, we interact, for example, gender with the treatment, the instrument, the enrollment control functions, as well as the segment. Table VI shows the results; the first two columns pertain to gender and the last three pertain to parental income.

31. With 25 children in the class, the per pupil cost equals $\frac{w}{25}$ (where w is the teacher wage), with 20 children in the class the per pupil cost equals $\frac{w}{20}$. The difference in per pupil cost is equal to $\frac{5}{100}w - \frac{4}{100}w = \frac{1}{100}w$.

32. Notice that by making the assumption that the average teacher wage equals the average future wages of those subjected to the policy, we abstract from productivity growth. This contributes to a downward bias in our rate of return calculations.

TABLE VI

HETEROGENEOUS EFFECTS OF CLASS SIZE

Dependent variable [# individuals]	Gender		Parents' income		
	Main effect (men)	Interact.	Interact. (1st Q)	Main effect	Interact. (4th Q)
Ability measures					
Cognitive ability, age 13	−0.0440**	0.027	0.0057	−0.0335**	0.0061
[$N = 5{,}116$]	(0.0188)	(0.020)	(0.0220)	(0.0163)	(0.0197)
Noncognitive ability, age 13	−0.0323**	0.002	0.0334	−0.0149	−0.0604***
[$N = 4{,}681$]	(0.0144)	(0.023)	(0.0270)	(0.0142)	(0.0227)
Academic achievement, age 16	−0.0330**	0.011	−0.0314	−0.0138	−0.0080
[$N = 5{,}318$]	(0.0155)	(0.024)	(0.0287)	(0.0152)	(0.0327)
Cognitive ability, age 18	−0.0242	—	—	—	—
[$N = 2{,}455$]	(0.0181)	—	—	—	—
Noncognitive ability, age 18	−0.0301**	—	—	—	—
[$N = 2{,}313$]	(0.0145)	—	—	—	—
Educational attainment (age 27–42)					
Years of schooling	−0.0554	0.0035	−0.0345	−0.0198	−0.0763
[$N = 5{,}588$]	(0.0341)	(0.0411)	(0.0692)	(0.0383)	(0.0663)
P(bachelor's degree)	−0.0065	−0.0021	−0.0089	0.0024	−0.0296**
[$N = 5{,}920$]	(0.0055)	(0.0077)	(0.0108)	(0.0059)	(0.0122)
Labor market outcomes (age 27–42)					
Earnings (eff. relative to the avg.)	−0.0108	0.0114	−0.0158	0.0021	−0.0296
[$N = 5{,}920$]	(0.0129)	(0.0277)	(0.0136)	(0.0074)	(0.0200)
ln(wage)	−0.0102	0.0092	0.0006	−0.0028	−0.0142*
[$N = 3{,}185$]	(0.0071)	(0.0127)	(0.0063)	(0.0046)	(0.0081)
P(earnings >0)	0.0023	−0.0077	−0.0259***	0.0055*	−0.0015
[$N = 5{,}920$]	(0.0037)	(0.0062)	(0.0080)	(0.0030)	(0.0079)

Notes. The estimates are based on representative samples of individuals born in 1967, 1972, 1977 or 1982 in one-school districts. All ability measures are standardized. The educational outcomes are measured in 2009, while the labor market outcomes have been averaged over the 2007–2009 period. Earnings effects (and their standard errors) are divided by average earnings level to facilitate interpretation. The ln(wage) estimates are restricted to wage-earners. Average class size in grades 4–6 is instrumented with Above threshold (=1 if school district enrollment in fourth grade exceeds the class size rule threshold in the enrollment segment). All models include the following controls for school district enrollment in grade 4: fixed effects for enrollment segment; linear controls for enrollment which are interacted with threshold and segment. In addition all models include the following baseline controls: municipality-by-cohort fixed effects, gender, dummy variables for month of birth, dummy variables for mother's and father's educational attainment, parental income, mother's age at child's birth, indicators for being a first- or second-generation Nordic immigrant, indicators for being a first- or second-generation non-Nordic immigrant, an indicator for having separated parents, and the number of siblings. Standard errors adjusted for clustering by enrollment count (77 clusters) are in parentheses. Asterisks indicate that the estimates are significantly different from zero at the ***1% level, **5% level, and *10% level.

The prime reason for showing heterogeneous impacts by gender is that we can estimate how cognitive and noncognitive abilities at age 18 are affected by variations in class size for men. We find that both estimates are negative, and that the effect on noncognitive ability is statistically significant. Taken at face value, the estimate for cognitive ability implies that 55% of the initial effect remains five years after the intervention ended and that 93% of the initial effect persists for noncognitive ability. Apart from that, the main message of the first two columns is that there are no significant gender differences in the effects of class size.

The last three columns of Table VI show the results for the parental income interactions, where parental income is the sum of both parents' average nonzero earnings (see the Online Appendix for further details). We construct two sets of interactions: one for the bottom 25% of the parental income distribution and one for the top 25%. The main effects thus pertain to individuals whose parents belong to the second and third quartiles of the income distribution, and the interaction effects are relative to this group.

With respect to parental income we find some differences, particularly in the longer run. The upper part of the table shows that there are no differential effects on cognitive ability in the short and medium run. The effect on noncognitive ability is, however, larger at the higher end of the parental income distribution.

The short-run impacts manifest themselves in two different ways on the labor market. For those from the lower end of the parental income distribution, a reduction of class size by one pupil improves the employment probability by 2 percentage points but has no effect on wages. For those from the high end of the parental income distribution, a reduction in class size yields no effect on the employment probability, but the wage effect is larger than in the rest of the distribution.

The differential labor market responses to variations in class size are interesting. We think that Swedish wage-setting institutions contribute to these findings. At the lower end of the wage distribution (where individuals from low-income families are more likely to end up), wages are typically determined by collective bargaining, while at the higher end of the wage distribution individual wage bargaining is more common (National Mediation Office 2011). Thus, given this bargaining structure, it is likely that interventions that affect individual skills have a wage effect at the higher end of the distribution; at the lower end, on the other hand, such interventions matter for the employment probability.

VII. CONCLUSION

This is the first article that documents significantly negative effects of class size in primary school on adult wages and earnings using a credible identification strategy. Previous attempts have been plagued by lack of precision (Chetty et al. 2011), unavailability of directly linked data on labor market outcomes (Krueger 2003; Schanzenbach 2007), or strong identifying assumptions (Dearden, Ferri, and Meghir 2002; Dustmann, Rajah, and van Soest 2003).

The effects are remarkably systematic across outcomes observed at various ages. Evaluated at the class size reduction in STAR (seven students), our estimates suggests improvements in cognitive ability at age 13 by 0.23 of a standard deviation and in pupil achievement at age 16 by 0.16 of a standard deviation, as well as an increase in adult wages by 4.4%. The wage impact corresponds to an increase of 0.16 of a standard deviation. The short-run effects thus tend to be highly persistent.

Our estimates of the wage effects of class size are substantially larger than imputations based on combining the short-run effect on cognitive ability and the association between cognitive ability and adult wages. The likely reason for this result is that observed cognitive ability captures limited dimensions of the skills that are priced on the labor market.

The wage effects are substantive, and given that we measure wages at age 27–42 these effects can arguably be considered permanent. Using the wage effects in a cost-benefit analysis reveals that the present value of the benefits outweigh the directly incurred costs. The internal rate of return is almost 18%. Moreover, reducing class size is a worthwhile investment even if we double the costs and cut the benefits in half.

Many previous studies have found negative effects of class size in primary school on short-term achievement. None of these studies has been able to demonstrate that these effects may have long-lasting effects on wages. There is no reason to believe that the permanence of the impact of class size is attributable to specificities of the Swedish context. There is, for instance, no strong correlation in class size across the stages of compulsory school in Sweden. There is also no evidence that the return to skill is higher in Sweden than elsewhere.[33]

33. Using data from the International Adult Literacy Survey, Leuven, Oosterbeek, and van Ophem (2004) estimate wage regressions for 15 different countries. In a specification with only years of education and experience (squared)

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at QJE online(qje.oxfordjournals.org).

APPENDIX TABLE A.1

DESCRIPTIVE STATISTICS, 1967–1982 BIRTH COHORTS

Variable [# individuals]	One-school districts	Full sample
Female	0.495	0.488
[$N = 5{,}920$; $N = 29{,}371$]	(0.500)	(0.500)
Mother's years of education	11.226	11.000
[$N = 5{,}920$; $N = 29{,}371$]	(2.761)	(2.708)
Father's years of education	11.096	10.743
[$N = 5{,}920$; $N = 29{,}371$]	(3.078)	(2.982)
Parental income	476,268	456,418
[$N = 5{,}920$; $N = 29{,}371$]	(232,763)	(204,826)
Cognitive ability, age 13	0.009	0.002
[$N = 5{,}116$; $N = 25{,}856$]	(1.022)	(1.001)
Noncognitive ability, age 13	0.028	0.011
[$N = 4{,}681$; $N = 23{,}864$]	(1.006)	(0.998)
Academic achievement, age 16	0.021	0.003
[$N = 5{,}755$; $N = 28{,}610$]	(1.015)	(1.001)
Cognitive ability, age 18 (men only)	0.048	0.004
[$N = 2{,}455$; $N = 12{,}949$]	(1.010)	(0.999)
Noncognitive ability, age 18 (men only)	−0.042	0.004
[$N = 2{,}313$; $N = 12{,}184$]	(0.986)	(1.001)
Years of schooling, age 27–42	13.519	13.491
[$N = 5{,}588$; $N = 27{,}771$]	(2.614)	(2.608)
Bachelor's degree, age 27–42	0.272	0.269
[$N = 5{,}920$; $N = 29{,}371$]	(0.445)	(0.443)
Earnings, age 27–42	232,248	242,372
[$N = 5{,}920$; $N = 29{,}371$]	(176,675)	(179,520)
P(earnings > 0), age 27–42	0.906	0.911
[$N = 5{,}920$; $N = 29{,}371$]	(0.292)	(0.285)
ln(wage), age 27–42	10.148	10.156
[$N = 3{,}185$; $N = 16{,}283$]	(0.279)	(0.274)

(continued)

the return to education in Sweden is 0.034 which is lower than in any of the other 14 countries. Including a measure of cognitive skill lowers the return to education in Sweden to 0.028, again lower than in any other country. The return to cognitive skill for Sweden is very close to the average of the 15 countries.

APPENDIX TABLE A.1

(CONTINUED)

Variable [# individuals]	One-school districts	Full sample
Class variables		
Class size in grade 4	24.337	23.329
	(3.843)	(3.437)
Class size in grade $4 > 30$	0.024	0.026
	(0.112)	(0.115)
Average class size grades 4–6	24.357	24.066
	(3.489)	(3.990)
School district variables		
Enrollment fourth grade	63.457	105.985
	(23.436)	(38.807)
Above class size rule threshold	0.408	0.470
	(0.492)	(0.499)
N individuals	5,920	29,371
N schools	191	1,129
N school districts	191	697
N clusters (enrollment counts)	77	165

Notes. The data are based on representative samples of individuals born in 1967, 1972, 1977 or 1982. All measures of cognitive ability, noncognitive ability and academic achievement have been standardized in the full sample. The educational outcomes are measured in 2009, while the labor market outcomes have been averaged over the 2007–2009 period. Wages are restricted to wage-earners. Standard deviations are in parentheses.

STOCKHOLM UNIVERSITY, INSTITUTE FOR EVALUATION OF LABOUR MARKET AND EDUCATION POLICY (IFAU), IZA, AND UPPSALA CENTER FOR LABOR STUDIES (UCLS)
IFAU, UPPSALA UNIVERSITY, AND UCLS
UNIVERSITY OF AMSTERDAM

REFERENCES

Angrist, Joshua D., and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114 (1999), 533– 575.

Bingley, Paul, Vibeke Myrup Jensen, and Ian Walker, "The Effects of School Class Size on Length of Post-Compulsory Education: Some Cost-Benefit Analysis," IZA Discussion Paper 1603, 2005.

Björklund, Anders, Melissa A. Clark, Per-Anders Edin, Peter Fredriksson, and Alan B. Krueger *The Market Comes to Education in Sweden—An Evaluation of Sweden's Surprising School Reforms* (New York: Russell Sage Foundation, 2005).

Black, Dan A., Jeffrey A. Smith, Mark C. Berger, and Brett J. Noel, "Is the Threat of Reemployment Services More Effective than the Services Themselves?" *American Economic Review*, 93, no. 4 (2003), 1313–1327.

Böhlmark, Anders, and Matthew J. Lindquist, "Life-Cycle Variations in the Association between Current and Lifetime Income: Replication and Extension for Sweden," *Journal of Labor Economics*, 24 (2006), 879–896.
Card, David, and Alan B. Krueger, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100, no. 1 (1992), 1–40.
Carrington, William J., and Kenneth R. Troske, "On Measuring Segregation in Samples with Small Units," *Journal of Business & Economic Statistics*, 15, no. 4 (1997), 402–409.
Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126 (2011), 1593–1660.
Dearden, Lorraine, Javier Ferri, and Costas Meghir, "The Effect of School Quality on Educational Attainment and Wages," *Review of Economics and Statistics*, 84 (2002), 1–20.
Dee, Thomas, and Martin West, "The Non-Cognitive Returns to Class Size," NBER Working paper 13994, 2008.
Duncan, Otis D., and Beverley Duncan, "A Methodological Analysis of Segregation Indices," *American Sociological Review*, 20 (1955), 210–217.
Dustmann, Christian, Najma Rajah, and Arthur van Soest, "Class Size, Education, and Wages," *Economic Journal*, 113 (2003), F99–F120.
Du Rietz, Lars, Ulf P. Lundgren, and Olof Wennås. "Ansvarsfördelning och styrning på skolområdet," DsU 1987:1, Ministry of Education, Stockholm, 1987.
Dynarski, Susan, Joshua M. Hyman, and Diane Whitmore Schanzenbach, "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion," NBER Working Paper 17533, 2011.
Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek, "The Devil Is in the (Institutional) Detail: Sorting and the RD Design in a Public School System," Unpublished manuscript, Stockholm University, 2012.
Gary-Bobo, Robert J., and Mohamed Badrane Mahjoub, "Estimation of Class-Size Effects, Using Maimonides' Rule and Other Instruments: The Case of French Junior High Schools," CEPR Discussion Paper 5754, 2006.
Härnquist, Kjell "Evaluation through Follow-up," In *Seven Swedish Longitudinal Studies in the Behavioral Sciences*, ed. Jansson, C.-G. (Stockholm: Forskningsrådsnämnden, 2000).
Heckman, James, Anne Layne-Farrar, and Petra Todd, "Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings," *Review of Economics and Statistics*, 78, no. 4 (1996), 562–610.
Hoxby, Caroline M., "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 115, no. 4 (2000), 1239–1285.
Krueger, Alan B., "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, no. 2 (1999), 497–532.
———, "Economic Considerations and Class Size," *Economic Journal*, 113 (2003), F34–F63.
Krueger, Alan B., and Diane M. Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal*, 111 (2001), 1–28.
Lee, David S., and David Card, "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 142 (2008), 655–674.
Lee, David S., and Thomas Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48 (2010), 281–355.
Leuven, Edwin, Hessel Oosterbeek, and Marte Rønning, "Quasi-Experimental Estimates of the Effect of Class Size on Achievement in Norway," *Scandinavian Journal of Economics*, 110 (2008), 663–693.
Leuven, Edwin, Hessel Oosterbeek, and Hans van Ophem, "Explaining International Differences in Male Wage Inequality by Differences in Demand and Supply of Skill," *Economic Journal*, 144 (2004), 478–498.

Lindahl, Mikael, "Home versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement," *Scandinavian Journal of Economics*, 107, no. 2 (2005), 375–394.
Mårdberg, Bertil, and Berit Carlstedt, "Construct Validity of the Swedish Enlistment Battery," *Scandinavian Journal of Psychology*, 34 (1993), 353–362.
McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142 (2008), 698–714.
National Mediation Office, "Summary of the Annual Report for 2010," National Mediation Office, 2011.
Schanzenbach, Diane Whitmore, "What Have Researchers Learned from Project STAR?" *Brookings Papers on Education Policy*, 2006/2007 (2007), 205–228.
Urquiola, Miguel, "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia," *Review of Economics and Statistics*, 88, no. 1 (2006), 171–176.
Urquiola, Miguel, and Eric Verhoogen, "Class-Size Caps, Sorting, and the Regression-Discontinuity Design," *American Economic Review*, 99 (2009), 179–215.