THE EFFECTS OF CLASS SIZE ON STUDENT ACHIEVEMENT: NEW EVIDENCE FROM POPULATION VARIATION*

CAROLINE M. HOXBY

I identify the effects of class size on student achievement using longitudinal variation in the population associated with each grade in 649 elementary schools. I use variation in class size driven by idiosyncratic variation in the population. I also use discrete jumps in class size that occur when a small change in enrollment triggers a maximum or minimum class size rule. The estimates indicate that class size does not have a statistically significant effect on student achievement. I rule out even modest effects (2 to 4 percent of a standard deviation in scores for a 10 percent reduction in class size).

I. INTRODUCTION

Class size reduction is probably the most popular and most funded school improvement policy in the United States. In 1996 the California legislature dedicated one billion dollars per year to class size reduction. The 1999 federal budget contained 12 billion dollars (over seven years) for the same purpose. Class size reductions are enacted often because they are popular with nearly every constituency interested in schools. Parents like smaller classes because their personal experience suggests that they themselves give more to each child when they have fewer children to handle. Even if parents in a school disagree bitterly about educational methods, they can agree that class size reduction is good: smaller classes give teachers the opportunity to practice more of *each* parent's favored educational method. Teachers, teachers' unions, and administrators like smaller classes for the same reasons parents do, but they may also like smaller classes for reasons that spring from self-interest. Teachers may like

* The author gratefully acknowledges the support of National Science Foundation grant SBR 9511343. She gratefully acknowledges the Connecticut Department of Education, whose staff were extremely efficient and knowledgeable. Neither the National Science Foundation nor the Connecticut Department of Education is responsible for any statements or errors in the paper. The author gratefully acknowledges helpful ideas from the editors, the anonymous referees, Henry Farber, Richard Freeman, Edward Glaeser, Daniel Hamermesh, Eric Hanushek, Lawrence Katz, Kevin Lang, Robert Lalonde, Richard Murnane, Steven Rivkin, Douglas Staiger, and seminar participants at the National Bureau of Economic Research, the University of Rochester, the University of Toronto, the University of Wisconsin, and the meeting of the Association of Public Policy and Management. This research was ably assisted by Bridget Terry and Ilyana Kuziemko.

The Quarterly Journal of Economics, November 2000

smaller classes because they reduce the effort that they must expend in order to deliver instruction. Teachers' unions may like class size reductions because they increase the demand for teachers. Administrators may like class size reductions because they increase the size of their domain. As a result of the policy's popularity, the twentieth century has been a period of continuous decline in class size, to the point where American elementary schools had, on average, 18.6 students per teacher in the 1997–1998 school year.[1]

Nevertheless, there are both economic and empirical problems with class size reduction policies. On the economic front, class size is a primary example of the education production function fallacy. It is conventional to estimate the relationship between educational inputs (like class size) and outputs (achievement) and to call the relationship an "education production function." This nomenclature suggests that inputs translate systemically into achievement, as they do in the production functions of profit-maximizing firms. The analogy is a false one, however, because firms' production functions are not just a result of their ability to turn inputs into outputs. A firm's production function is the result of maximizing an objective (profits), given a production possibilities set. It is not obvious that schools have stringent achievement maximization objectives imposed on them. As described above, class size reductions can fulfill a variety of objectives, not all of which are related to achievement. Thus, while class size reduction always affords *opportunities* for increased investment in each child's learning, it is not obvious that every school takes up such opportunities. The actual effect of reducing class size will depend on the incentives a school faces. Put another way, if a policy-maker wants to predict the effect that a proposed class size reduction would have, she should rely on evidence from schools that face incentives that are similar to the incentives that schools would face under the proposed policy.

On the empirical front, class size is difficult to study.[2] The vast majority of variation in class size is the result of *choices* made by parents, schooling providers, or courts and legislatures. Thus,

[1] See National Center for Education Statistics [1999]. There are differences between the student-teacher ratio and class size, but the differences are less of a concern for elementary schools than for secondary schools. In any case, the differences are not relevant to the empirical work in this paper, because I use class size as reported by schools.

[2] Surveys of the evidence on class size include Hanushek [1996, 1986], Card and Krueger [1996], and Betts [1995].

most of the observed variation in class size is correlated with other determinants of student achievement and is likely to produce biased results. This may appear to be an obvious point, but though researchers often claim that the variation they use is not endogenous to student achievement, they rarely go on to explain where the variation *does* come from. The processes by which school inputs are determined should make us doubt that variation in school inputs is exogenous unless there is some explicit reason why we should think it is.

This criticism does not apply to explicit experiments that randomly assign some students to small classes and other students to large ones. Project Star is an experiment of this type, and evidence based on it has manifest advantages.[3] These advantages, however, are offset by a few disadvantages. Explicit experiments are rare (tempting interpreters to extrapolate the results unduly), many experiments take place in developing countries (so that the range of inputs is not relevant for the United States), and—most importantly—the actors in an experiment are aware of it. For instance, the schools in a class size experiment may realize that if the experiment fails to show that the policy is effective, the policy will never be broadly enacted. In such cases the schools have incentives that the fully enacted policy would not give. That is, the experiment alters the incentive conditions, so that the production function being estimated is not the production function that would be in effect if the policy were fully enacted. In addition, some individuals temporarily increase their productivity when they are being evaluated. This phenomenon, known as the "Hawthorne effect," can make policies appear to have productivity effects that they would not have if fully enacted. Finally, individuals sometimes try to undo the randomness of the experiment. For instance, some administrators may try to fill the small classes with children who are most in need of individual attention (generating results that are biased against finding that class size reduction works). Other administrators may assign their best teachers to the small classes or monitor the small classes more (generating results that are biased toward finding that class size reduction works).

In this study I attempt to address the empirical and economic problems with two identification strategies, both of which use variation in class size that comes from population variation. The

[3] Project Star was an explicit experiment in class size reduction in Tennessee elementary schools. See Krueger [1999] for a description of the Project Star results.

first strategy uses natural randomness in the population, and its logic is straightforward. Consider a school attendance area that has a population that is in steady state. There is still natural randomness in the timing of births such that the entering kindergarten cohort varies somewhat in size. This variation is not fully smoothed because there is discreteness in school entry rules (for instance, children born between January 1 and December 31 in year t must enroll in first grade in year $t + 6$) and because the number of classrooms in each school is an integer. If one thinks of a school with one classroom per grade, then natural randomness in the population translates directly into differences in class size between cohorts. For instance, suppose that a school attendance area has an unusually small number of five-year olds with birthdays in November and December 1985 but has the "deficit" made up by an unusually large number of five-year olds with birthdays in January and February 1986. These small timing differences would typically make for an unusually small kindergarten cohort (say, 15 students) in the 1990–1991 school year and an unusually large kindergarten cohort (say, 25 students) in the 1991–1992 school year. The first cohort might persistently experience small classes in grades kindergarten through 6, while the subsequent cohort might persistently experience large classes. Essentially, the two cohorts are randomly assigned different class sizes.

I implement the first identification strategy by isolating the random component of population variation using long panels of data on enrollment and kindergarten cohorts in Connecticut school districts. The long panels allow me to eliminate nearly all smooth changes in population. I use residuals that remain after fitting a quartic function of time *separately* for each grade in each school.

In the second strategy I use the fact that class size jumps abruptly when a class has to be added to or subtracted from a grade because enrollment has triggered a maximum or minimum class size rule. Returning to the previous example, suppose that the 1992–1993 kindergarten cohort were 26 students and the district's maximum kindergarten class size were 25. Then there would be two kindergarten classes of 13 students each in 1992–1993. Although the difference in cohort size between 1991–1992 and 1992–1993 would be only one student, the difference in class size would be twelve students. The logic of the identification strategy is that there is a discontinuous relationship between class size and enrollment at certain known levels of enrollment while there is a smooth relationship between achievement and the determinants of enrollment. I use the panels of data to observe small changes in enrollment associated with changes in the number of classes in each grade in each school. I use information on each district's class size rules to determine whether change in the number of classes was purely the result of the small change in enrollment triggering a rule. I implement the second identification strategy by comparing the class size and achievement of adjacent cohorts who immediately precede and succeed each such event.

The two identification strategies are independent of one another and provide a check on one another's results. I provide a number of other specification tests as well.

One nice consequence of using population variation is that the range of class size for which I obtain estimates is the range that is relevant for policy. Another nice consequence is that I observe schools functioning under the incentive conditions that they normally experience. The one disadvantage of natural population variation is that a teacher may adjust her teaching methods more over two or three years of small class size than she does over one year of small class size (even if she periodically experiences small classes). In Project Star, most of the effect of small class size occurred after one year, without teachers' being trained to alter their teaching methods. These facts suggest that teachers can adjust quickly and without special training, if they have an incentive to do so. In short, these facts suggest that the transitoriness of small class size due to population variation should not be a problem, but I discuss the issue carefully in interpreting my results.[4]

II. Sources of Variation in School Inputs and the Potential for Bias

Parents' choosing schools by choosing their residences is probably the single largest source of variation in school inputs. Between-district variation in school inputs generated by parents' choices is likely to generate upward biased estimates of the

[4] Population variation makes teachers experience small and large classes repeatedly, but not predictably. Note that class size is *not* transitory from the point of view of a cohort of students. A cohort in a school tends to experience either small or large classes consistently.

efficacy of inputs. The same may be said for systematic variation within a district over time. For instance, class size reductions will appear to be more efficacious than they really are if parents who contribute more to their children's learning also choose school districts that offer smaller class sizes. When we make simple comparisons of schools in cross-section data or time-series data, there is likely to be bias in favor of class size reductions.

If we identify parents who have similar attributes, there is ample but somewhat different potential for bias. Parents choose school inputs endogenously, based on their child's ability and prior achievement in school. These endogenous choices may be compensatory (greater inputs for children who exhibit poor achievement), reinforcing (greater inputs for gifted children), or both. Thus, when we compare students with similar families (using, say, cross-section data with extensive controls for family background), the sign of the bias is ambiguous.

Similarly, we cannot predict the sign of the bias generated by the choices of schooling providers, such as administrators and teachers. If providers attend more to the demands of parents who contribute more to their children's learning, inputs and parental contributions will be positively correlated, generating upward biased estimates of the efficacy of inputs. On the other hand, if providers attend more to children with learning problems, estimates will be biased downward.

The final players who determine school inputs are state and federal judges and legislators, who mandate and fund increased school inputs for certain students. Policy-makers pursue both compensatory and reinforcing policies, but they tend to devote the majority of the resources at their disposal to compensatory policies.[5] The negative bias resulting from the use of compensatory policies, however, is often offset by positive omitted variables bias caused by policy-makers' simultaneous pursuit of complementary policies. For example, policies that decreased racial discrimination in school inputs were implemented simultaneously with policies that decreased racial discrimination in employment. Both types of policies could lead minority students to have higher achievement.

[5] See Salmon et al. [1995] for evidence on the prevalence of compensatory policies in state school finance. More than 80 percent of federal money for elementary and secondary education is devoted to compensatory policies: Title I, bilingual education, special education, and the free and reduced-price lunch program.

In short, it is not surprising that empirical results differ (that is, suffer from different biases) depending on the source of variation in school inputs that they use.

There is a difference between variation that is not obviously biased and variation that has an explicit reason to be random. The systematic links between school inputs and other determinants of student outcomes may be *obscure* without the variation in inputs being *exogenous*. Explicitly articulating a source of exogenous variation is preferable to simply eliminating all apparent sources of bias. This is what I attempt to do in this paper.

III. Empirical Strategy

Consider the achievement of students in grade i of school j in district k in cohort t. It is determined by class size as well as unobserved attributes like student ability and parental contributions to learning. A general "education production function" that subsumes most common specifications is

$$(1) \qquad A_{ijkt} = \beta_1 \log (C_{ijkt}) + \mathbf{I}_t\beta_2 + \mathbf{I}_j\beta_3 + \mathbf{X}_{ijkt}\beta_4 + \epsilon_{ijkt},$$

where A_{ijkt} is achievement; $\log (C_{ijkt})$ is the natural log of class size; $\mathbf{I}_t$ is a vector of cohort indicator variables; $\mathbf{I}_j$ is a vector of school indicator variables; $\mathbf{X}_{ijkt}$ is a vector of observed student, parent, and community characteristics; and ϵ_{ijkt} is all other determinants of achievement, including the unobserved attributes of the students, parents, and community. Few studies actually include all of the terms in equation (1), but most studies include some subset of them.

If the measure of achievement is a test score, it is often divided by the standard deviation of students' scores on the test. This common practice (which I follow in this paper) facilitates understanding of unfamiliar test scores and allows comparisons to be made across studies that use different tests. It is now common to use the natural log of class size to take account of the fact that a one-student reduction is proportionately larger from a base of 17 students, say, than from a base of 35 students. The vector of cohort indicator variables is included to allow for tests that change slightly from year to year and to allow for teachers who adjust their teaching to a test's content. For instance, if a test were slightly easier for all fourth graders in 1996 than it was for all fourth graders in 1995, the slight easing would be picked up by the cohort indicators. The vector of school indicator variables is included to control for any attributes of schools' populations that are constant across time—especially relatively unchanging attributes of the community in which the school is located. The vector **X** typically includes variables that describe the racial composition and free lunch eligibility of students. It also includes variables that describe the racial composition, educational attainment, and income of local households (although school fixed effects, if included, absorb any such variables that are constant across cohorts of students).

1. The First Identification Method

By definition, the log of average class size is equal to log (E) − log (n), where E is regular enrollment, and n is the number of classes. For now, let us focus on enrollment and suppose that the number of classes is fixed for a given grade in a given school. I return to the changes in the number of classes below.

Enrollment is a function of student, parent, and community characteristics (observed and unobserved). In addition, there is random variation in the population of children who are in the age range appropriate for a given grade in a given year. That is, actual enrollment has a deterministic component, $\check{E}(\mathbf{X},\epsilon)$, which is what enrollment would be if the timing and number of births were a deterministic function of the population's observed and unobserved characteristics; and it has a random component u, which is the variation in enrollment that results from the fact that biology causes random variation in the timing and number of births. One expects that u affects E proportionally, so one can write[6]

$$(2) \qquad E_{ijkt} = \check{E}_{ijkt}(\mathbf{X}_{ijkt},\epsilon_{ijkt}) \cdot u_{ijkt},$$

$$\text{or} \qquad \log (E_{ijkt}) = \log (\check{E}_{ijkt}(\mathbf{X}_{ijkt},\epsilon_{ijkt})) + \log (u_{ijkt}).$$

Log (u) is not correlated with X and ϵ, which are determinants of achievement, but log (u) is a determinant of log (E), so a consistent estimate of log (u) is a good instrument for class size. That is, if an estimate of log (u) is consistent, then it fulfills the two basic instrumental variables conditions: it is correlated with log (E) and uncorrelated with ϵ. I attempt to get a consistent estimate of log (u) using the fact that the deterministic part of enrollment changes much more smoothly than actual enrollment

[6] It is natural to suppose that the share, not the absolute number, of "deviantly" timed births is constant across populations of different sizes.

does for a particular grade in a particular school in a particular year. Consider a school attendance area that has, for various reasons, a positive trend in the number of households with school-aged children. (The trend could be nonlinear.) The deterministic part of enrollment in each grade would be a relatively smooth function of the trend. But actual enrollment in each grade would deviate from this relatively smooth function and might not even be a monotonic transformation of the trend in the number of households with school-aged children. Recall the example of the school attendance area that had an unusually small number of children with birthdays in November and December 1985 and an unusually large number with birthdays in January and February 1986. These timing differences would typically generate a small kindergarten cohort in 1990–1991 and a large one in 1991–1992, and the later cohort could experience larger class sizes *even if* the school attendance area had a negative trend in the number of households with school-aged children. Moreover, if one were to de-trend enrollment, one would find that the 1990–1991 cohort had a negative residual and the 1991–1992 cohort had a positive residual.

Any log ($\check{E}$) that changes smoothly over time can be approximated by a grade-school-specific intercept and a grade-school-specific polynomial in time. That is, we can write

$$(3)\quad \log(\check{E}_{ijkt}) = \alpha 0_{ijk} + \alpha 1_{ijk} t + \alpha 2_{ijk} t^2 + \alpha 3_{ijk} t^3 + \alpha 4_{ijk} t^4 + \cdots,$$

or $\log(E_{ijkt})$

$$= \alpha 0_{ijk} + \alpha 1_{ijk} t + \alpha 2_{ijk} t^2 + \alpha 3_{ijk} t^3 + \alpha 4_{ijk} t^4 + \cdots + \log(u_{ijkt}).$$

I estimate such an equation separately for *each* grade in *each* school. I typically have 24 years of enrollment data for each regression. I show results that use up to a quartic in time because quartics appear to capture all of the smooth variation over time in enrollment within a grade within a district.[7] The estimated residual should be a consistent estimate of log (u), which is the instrument we need for class size.

In short, the first identification strategy has three intuitive steps: one, obtain estimates of the random part of enrollment variation; two, use the random variation in enrollment to identify random variation in class size; three, see how achievement is

[7] In fact, the estimated residuals hardly change in the move from a cubic to a quartic, quintic, or sixth-order polynomial.

affected by random variation in class size. Formally, the first identification strategy requires the following procedure. First, estimate equation (3) separately for each grade in each school and obtain the estimated residuals. Stack the estimated residuals to get a vector of the estimated residuals for each school: log $(\hat{u}_{ijkt})$. Second, estimate the following first-stage equation for each grade:

$$\log (C_{ijkt}) = \delta_1 \log (\hat{u}_{ijkt}) + \mathbf{I}_t\delta_2 + \mathbf{I}_j\delta_3 + \mathbf{X}_{ijkt}\delta_4 + \nu_{ijkt}, \tag{4}$$

and obtain predicted log (C_{ijkt}). Third, estimate equation (1) by Two Stage Least Squares (2SLS), using predicted log (C_{ijkt}). Calculate correct standard errors for the 2SLS procedure.[8] Notice that the procedure uses *within-school* comparisons of enrollment, class size, and achievement. A school fixed effect is taken out of enrollment to form log $(\hat{u}_{ijkt})$; a school fixed effect is estimated in the first-stage equation; and a school fixed effect is estimated in the second-stage equation. One cohort in a school is being compared with others in the same school, where the difference between the cohorts is that one is larger than the others due to (what appear to be) purely random circumstances.

The method just described exploits the fact that aggregate characteristics that affect achievement, $\mathbf{X}$ and ϵ, change much more continuously than enrollment in a specific grade-school-time does. Yet, because parents can respond directly to the class size they observe their child experiencing, the method leaves open a small route for bias. Consider a parent who observes that his child's class is unusually large. Even if the cause of the large class is random population variation, the parent might decide to have his child transferred to another school in the same district, might move to another district, might send his child to a private school, or might attempt to have his child held back a grade or advanced a grade. Such reactions, although probably rare, would have the potential to make $\mathbf{X}$ and ϵ endogenous to u. A parent who would react this way would have to be unusually concerned about education, able to pay for a move, able to pay for private schooling, or able to convince school administrators to allow a transfer. One expects that such a parent would, in any case, make an unusually large contribution to his child's education, so that the endogeneity under consideration would probably make us overestimate the efficacy of class size reductions. In other words, classes that were

[8] That is, calculate the standard errors using the actual data on class size, not the predicted data.

randomly large would end up with a disproportionately small share of education-concerned parents. Fortunately, one can do better than speculate about the size and sign of this bias: a simple modification of the estimation method eliminates the problem.

Rather than carry out the instrumental variables procedure at the school level, one can aggregate equations (1) and (3) to the district level and carry out the procedure at the higher level of aggregation. At the district level, transfers among schools within the district will cancel out, so residuals from the district-level version of equation (3)—

$$(5)\quad \log(\overline{\mathbf{E}}_{ikt}) = \tilde{\alpha}0_{ik} + \tilde{\alpha}1_{ik}t + \tilde{\alpha}2_{ik}t^2 + \tilde{\alpha}3_{ik}t^3 + \tilde{\alpha}4_{ik}t^4 + \cdots + \log(u_{ikt})$$

—give us a credibly consistent estimator for log (u) that has no potential to be correlated with $\mathbf{X}$ or ϵ through parents' reacting to large class size by transferring a child to another school within the district. $\overline{\mathbf{E}}_{ikt}$ is enrollment in grade i in district k for cohort t, summed over all of the schools in the district. Carrying out the procedure at the district level eliminates bias caused by transfers; it also has advantages because more years of achievement data are available at the district level. On the other hand, carrying out the procedure at the district level reduces the explanatory power of the procedure. In particular, the explanatory power contributed by large school districts is reduced because random population variation averages out to a great extent within each cohort over a large district. (Elementary schools in large districts, however, are small enough that large districts do contribute significantly in school-level estimation.)

The district-level procedure does not entirely eliminate the potential for bias caused by parents' reacting to class size. Parents could still shift their child to a private school, have their child held back or advanced a grade, or move out of the district once they observed that their child's class was going to be unusually large. Fortunately, one can remove this potential for bias by using data on the number of children in each district who were age five at the school entry date. In other words, one can observe the potential kindergarten cohort at the district level ("$K5$") and use it as the source of random variation in class size. One simply estimates a version of equation (3) with the potential kindergarten cohort as the dependent variable:

$$\log(\bar{K}5_{ikt}) = \check{\alpha}0_{ik} + \check{\alpha}1_{ik}t + \check{\alpha}2_{ik}t^2 + \check{\alpha}3_{ik}t^3 + \check{\alpha}4\ _{ik}t^4 + \ldots + \log(u_{ikt}). \tag{6}$$

Equation (6) gives us a credibly consistent estimator for log (u) that has no potential to be correlated with $\mathbf{X}$ or ϵ through parents reacting to idiosyncratically large class size by moving to another district, sending a child to private school, or shifting a child to a different grade. In addition to the disadvantages discussed above for district-level estimation, the disadvantage of using kindergarten cohort residuals is that they will be stronger instruments for class size in early elementary grades than in later elementary grades because exogenous student mobility weakens the correlation between kindergarten cohort size and later grades' cohort sizes.

Thus far, I have not discussed changes in the number of classes n in a grade in a school. My second identification method exploits these changes, but they are simply a nuisance for my first identification method. The costs and benefits of adding another class depend not only on how much local parents care about schooling but also on actual enrollment in any given year (even if the rise or shortfall in actual enrollment comes from random variation). Thus, if one carries out the procedure for the first identification method and ignores changes in the number of classes, the monotonicity condition for instrumental variables will occasionally be violated: an increase in enrollment will *reduce* class size if it triggers an increase in the number of classes.[9] There is a simple way to adjust the first identification method so that the monotonicity condition is never violated. The procedure described above is valid so long as the variation in enrollment does not trigger a change in the number of classes. Therefore, I use variation in enrollment that is not just within-school but is within an expected number of classes. In other words, instead of having school indicator variables in the first- and second-stage equations, there is an indicator variable for each combination of a school and expected number of classes. That is, there is a vector of indicator variables for combinations like the following: the school is j and its second grade is expected to have two classes, the school is j and its second grade is expected to have three classes, and so on. The

[9] See Angrist, Imbens, and Rubin [1996] for a discussion of the monotonicity condition for instrumental variables.

logic is straightforward. If enrollment in a school's second grade is randomly higher this year than it was last year but is such that there are two second grade classes in both years, then the regression compares the difference in achievement between the two years with the difference in class size. If enrollment in a school's second grade is randomly higher this year and it triggers a maximum class size rule so that this year there are *three* second grade classes, then the regression does not compare the two years. Notice that the *expected* number of classes is what matters. I use districts' maximum and minimum class size rules to determine when an enrollment change would be expected to trigger a change in the number of classes, since it is at these times that the monotonicity condition would be violated. (If a school changes the number of classes for reasons unrelated to enrollment but related to, say, changes in parents' preferences, the monotonicity condition is not violated.) Calculation of the expected number of classes is discussed in the next subsection.

Summing up, the first identification strategy proceeds as follows. First, estimate equation (3) separately for each grade in each school, and obtain the estimated residuals, log $(\hat{u}_{ijkt})$. Second, estimate the following first-stage equation, in which there is a fixed effect for each school-expected number of classes combination:

$$\log (C_{ijkt}) = \delta_1 \log (\hat{u}_{ijkt}) + \mathbf{I}_t\delta_2 + \mathbf{I}_{j,n_j}\delta_3 + \mathbf{X}_{ijkt}\delta_4 + \nu_{ijkt}. \tag{7}$$

$\mathbf{I}_{j,nj}$ is vector of indicator variables for combinations of schools and expected number of classes. Third, estimate the following achievement equation, in which there is a fixed effect for each school-expected number of classes combination:

$$A_{ijkt} = \beta_1 \log (C_{ijkt}) + \mathbf{I}_t\beta_2 + \mathbf{I}_{j,n_j}\beta_3 + X_{ijkt}\beta_4 + \epsilon_{ijkt}. \tag{8}$$

Calculate correct standard errors for the 2SLS procedure. Repeat the procedure with district-level enrollment and with district-level kindergarten cohorts.

2. The Second Identification Method

The second identification method does not treat changes in the number of classes as a nuisance; it exploits them. It makes use of the fact that changes in the number of classes in a grade can produce abrupt changes in class size. The simplest way to use these discontinuities is the cross-section method of exploiting maximum class size thresholds. Angrist and Lavy [1999] illustrate this method using Israeli schools. (Israeli schools have a maximum class size of 40; most American districts have maximum class sizes in the range of 20 to 30 students.) For instance, if a school has a maximum class size threshold of 25, it puts students into one class until enrollment is 25, puts students into two classes until enrollment is 50, and so on. Its rule can be written as

$$C_{ijkt} = \frac{E_{ijkt}}{\text{int}\,[(E_{ijkt} - 1)/C^{\max}] + 1}, \tag{9}$$

where $C^{\max}$ is 25 and int (z) is the greatest integer less than or equal to z. Class size varies abruptly and predictably when enrollment is at a multiple of 25. These discontinuities provide identification because the difference in the underlying population that produces enrollment of 25 versus 26 is very small (and should have a correspondingly small effect on achievement), but the difference in class size for enrollment of 25 versus 26 is large (and should have a significant effect on achievement if reductions in class size are efficacious). Thus, the change in the predicted class size between enrollment of 25 and enrollment of 26 based solely on the rule given by equation (9) is a good instrument for the actual difference in class sizes between schools with enrollment of 25 and 26. The same is true for 50 and 51, 75 and 76, and so on.

There are three essential things to understand about this method of identification. First, the identification is independent of the identification that comes from using log (u) as an instrument for class size, so the two methods can be used as checks on one another.

Second, between the discontinuities, predicted class size varies with actual enrollment, which is, of course, a function of $\mathbf{X}$ and ϵ. Therefore, predicted class size is *not* a valid instrument *except* when the rule triggers a change in the number of classes. Put another way, the estimates will be consistent only if identification relies *solely* on the discontinuities in equation (9). All variation in predicted class size that is not generated by a rule-triggered change in the number of classes is suspect and must be discarded if bias is to be eliminated. In cross-section data one does not observe actual changes in the number of classes, so the only nonsuspect variation is the variation at multiples of maximum class size—the difference in achievement for enrollment of 25 versus 26, for enrollment of 50 versus 51, et cetera. In cross section data other variation in enrollment is suspect because it is likely to be between-district variation that reflects differences in the underlying populations (**X** and ϵ) and could even be endogenous to realizations of class size. Some schools *routinely* have larger class sizes than others because of the way the rules function, and parents can endogenously choose schools taking realized class size into account. Discarding all suspect observations, however, places great demands on cross-section data, since the results will depend on there being sufficient occurrences of enrollment at multiples of maximum class size. Angrist and Lavy [1999], for instance, are able to do only some of the desirable discarding because their cross-section data contain too few occurrences of enrollment in the right ranges. Below, I present cross-section results that demonstrate what happens as one discards more and more of the suspect observations. Since my data are actually panel data, I am able to employ a within-district method (described below) that is more powerful and less subject to bias than the cross-section method.

Third, identification arises only when the rule binds, so if one uses a rule that binds only in some schools, one learns about the effects of class size only for those schools. For instance, in Angrist and Lavy's [1999] data, the maximum class size rule does not bind in districts that serve well-off households. It is useful to estimate the effect of class size only for less-well-off students, but one must be careful to interpret the results appropriately. If better-off districts actually have maximum class size rules of their own that they follow, then using a statewide rule that does not bind everywhere is throwing away useful variation. Since there is typically not much useful variation anyway for discontinuity-based identification strategies, it is important to use all that exists.

Given these issues about identification based on discontinuities, I use changes in the number of classes that are generated by small *within-school* changes in enrollment that trigger a *district's* maximum or minimum class size rule. This method is more accurate and less prone to bias than the cross-section method because one can follow enrollment in a grade in a school over time and actually see every occasion on which the rules are triggered by small changes in enrollment. The method also produces more precise estimates because it compares adjacent cohorts within a school, who are likely to be similar *except* for their different class size experiences. Finally, this method has the advantage that it uses variation from all sorts of districts. Districts choose rules that are relevant for them, and as long as the rules are stable, they generate useful discontinuities in class size. The within-school regression discontinuity method requires district-by-district information on class size rules, which is onerous to collect. I obtained information on each district's rules by surveying superintendents (see below). Note that, as long as each district's rule is stable over the period in question, the rules could be endogenous to the underlying characteristics of the districts and the second identification method would still produce consistent results. This is because the second identification method relies on *within-district* variation in class size.[10]

The second identification method ("within-school regression discontinuity") has a very simple procedure. First, I identify all of the events in which a school increased or decreased the number of classes in one of its grades. Second, within this group I identify all the events in which the change in the number of classes was predictable, given just the change in enrollment and the district's maximum and minimum class size rules. I keep this subset (which is, in practice, 78 percent of all events where the number of classes changes). That is, the expected number of classes is given by

$$E(n_{ijkt}) = n_{ijk,t-1} + \mathbf{I}_{ijkt}^{\text{add}} - \mathbf{I}_{ijkt}^{\text{subtract}}, \tag{10}$$

where

$$\mathbf{I}_{ijkt}^{\text{add}} = 1 \qquad \text{if } \frac{E_{ijkt}}{n_{ijk,t-1}} > C^{\max}; \qquad 0 \text{ otherwise,}$$

and

$$\mathbf{I}_{ijkt}^{\text{subtract}} = 1 \qquad \text{if } \frac{E_{ijkt}}{n_{ijk,t-1}} < C^{\min}; \qquad 0 \text{ otherwise.}$$

I keep the subset of events where the number of classes actually increased and $\mathbf{I}_{ijkt}^{\text{add}} = 1$ and where the number of classes actually

[10] There is a caveat to this statement. Since districts can set lower or higher maximum class sizes, districts will generate useful class size variation in slightly different ranges. For instance, one district's useful variation in class size may tend to be in the range from 16 students to 25 students, while another's may tend to be in range from 18 to 27 students. If one were to find that a reduction in class size was, say, more efficacious when it occurred above some class size, then one would be unsure whether the greater efficacy was due to decreasing returns to reductions in class size or greater efficacy in the sort of schools that typically choose higher maximum class size. One could then try to sort out the explanations by examining the characteristics of districts with lower and higher maximum class sizes. This problem does not arise, in practice, in this study.

decreased and $\mathbf{I}_{ijkt}^{\text{subtract}} = 1$. Third, within the subset I keep the events in which the change in enrollment that triggered the change in class size was smaller than 20 percent. For instance, if enrollment rose from 40 to 48, and it triggered a change in the number of classes, I keep the event. However, if enrollment rose from 40 to 54, I discard the event. The reason is that regression discontinuity methods depend on a *modest* change in a continuous variable, like enrollment, triggering a *large* change in a discrete variable, like the number of classes. If some change in the underlying circumstances of a school were to make both enrollment and the number of classes jump by a large amount, the event would be inappropriate for regression discontinuity methods. In practice, I keep 94 percent of the subset at this stage.

Having identified a set of events where the number of classes changes because a modest change in enrollment triggers a maximum or minimum class size rule, I estimate a first-differenced version of the achievement equation—

$$\text{(11)}\quad [A_{ijkt} - A_{ijk,t-1}] = \beta_1[\log(C_{ijkt}) - \log(C_{ijk,t-1})] + \mathbf{I}_t\beta_2 + [\mathbf{X}_{ijkt} - \mathbf{X}_{ijk,t-1}]\beta_4 + [\epsilon_{ijkt} - \epsilon_{ijk,t-1}]$$

—using just the cohorts immediately before and after each event. Intuitively, if school *j*'s third grade enrollment is modestly higher this year than it was last year, and the enrollment increase triggers a maximum class rule so that the number of classes rises and class size falls, then I compare the achievement of this year's third grade cohort with that of last year's third grade cohort.[11]

[11] One may worry about "nonevents"—occasions on which a small change in enrollment should have triggered a change in the number of classes but did not. It turns out that only 9 percent of would-be trigger events are actually not associated with a change in the number of classes. Moreover, discussions with superintendents suggest that alert parents tend to make sure that maximum class size rules are enforced but try to prevent the enforcement of minimum class size rules. Thus, ignoring nonevents may produce a small bias in favor of class size being efficacious (smaller class sizes are associated with more alert parents). Such bias is not a concern, given the results.

One can carry out a district-level version of the procedure for the second identification method. The district-level procedure eliminates the possibility that the results are due to parents' responding (to large class sizes) by transferring their children to other schools within the school district. Relative to the school-level procedure, the district-level procedure has all the same advantages and disadvantages as it has in the first identification method. Rather than carry out the district-level procedure for the second identification method, I simply examined each event to determine whether there were offsetting enrollment changes in other schools in the same district. I did not find any such offsetting changes. Event-by-event examination is possible because the number of events is limited.

3. A Note on Single-Year versus Multiple-Year Effects of a Change in Class Size

So far, I have written all of the equations as though a change in class size this year generates a change in achievement by the end of the year. This is because recent empirical results suggest that such single-year effects are the important effects (see Krueger [1999] and Angrist and Lavy [1999]). It may be, however, that such specifications are not a fair test of class size because a student needs to be in smaller classes for a few grades before there is any effect. In the empirical work that follows, I do provide results for class size in the most recent year, but I mainly focus on specifications that use the average class size that a cohort has experienced up until the time it takes the test.

I focus on the specifications that use average class size because they favor finding that class size is efficacious. The reason is that a cohort usually experiences small or large classes consistently,[12] so that almost none of the difference between the average class size experienced by one cohort and the next within a school is likely to be caused by measurement error. If there were measurement error, it would wash out once the class size experienced by a cohort was averaged over a few grades. It is important to remember that the identification strategies rely on *cohort-to-cohort* differences in class size and that each cohort experiences relatively unchanging class size.

IV. Data

The two identification methods create a number of data requirements. First, because the integer nature of teachers and classrooms is useful for making natural population variation translate into variation in class size and composition, data on the elementary grades is needed. Elementary classes are less divisible than secondary school classes because the standard method of elementary school instruction is one teacher spending the majority of each school-day with a regular group of students in one classroom.[13] Also, class size is well-defined in elementary schools but poorly defined in middle and high schools, where students may experience different class sizes in different subjects. The

[12] This point is demonstrated below.

[13] A class is the group of students who spend the majority of the school day with one teacher. The measure of class size excludes pull-out instruction by special education teachers or aides.

resulting emphasis on elementary class size fits the empirical and pedagogical debates, which have focused on class size in early grades. Another reason for focusing on elementary class size is that elementary schools are not large. In very large schools, natural population variation averages out to a great extent within each cohort.[14] Finally, since school cohorts are defined by birth date, one needs data on population-by-age at the school entry cutoff date (which is December 31 in Connecticut).[15]

Connecticut school data are particularly appropriate for the empirical strategy. The state has 649 elementary schools that belong to 146 elementary districts.[16] Overall, 25 percent of the schools have typical cohort sizes smaller than 46 students; 50 percent have smaller than 63 students; and 75 percent have smaller than 92 students.[17] Districts are essentially towns in Connecticut and, for many years, the towns collected annual Enumerations of Children (population-by-age data as of January 1 for all school-aged children). In the last few years, similar data have been compiled by Claritas Incorporated.[18] The Enumeration of Children and Claritas are the source of the potential kindergarten cohort data. Finally, every year since 1986, Connecticut has administered statewide tests in the fourth, sixth, and eighth grades.[19] From 1986 to 1991, test data are available by district. From 1992 onward, test data are available by school (as well as by district). I use six years of school-level test data (from 1992–1993 to 1997–1998)

[14] For the district-level versions of the procedures (which are essentially specification tests), it is useful to have some districts that are small (that contain only one to three elementary schools).

[15] In Connecticut a child is ordinarily enrolled in kindergarten if he will be five by December 31 of the school year.

[16] Elementary schools are schools that contain some combination of grades 1 to 6. Most elementary schools in Connecticut contain grades 1 through 6, but some districts have separate schools for the lower elementary grades and upper elementary grades. See notes to Table I.

[17] Among schools that are in districts with median household income below the twenty-fifth percentile for Connecticut, the distribution of cohort sizes is as follows: 43 = 25th percentile, 57 = 50th percentile, 78 = 75th percentile. Among schools that are in districts with median household income above the seventy-fifth percentile for Connecticut, the distribution of cohort sizes is as follows: 45 = 25th percentile, 66 = 50th percentile, 89 = 75th percentile. Among schools that are less than 5 percent African-American, the distribution of cohort sizes is as follows: 46 = 25th percentile, 63 = 50th percentile, 93 = 75th percentile. Among schools that are more than 10 percent African-American, the distribution of cohort sizes is as follows: 44 = 25th percentile, 61 = 50th percentile, 79 = 75th percentile. The last group of schools is almost exclusively urban.

[18] Some districts combine two small towns. In such cases, the towns' population-by-age statistics are aggregated.

[19] Between 1979 and 1985, Connecticut administered statewide achievement tests in the ninth grade. Ninth grade scores are not ideal for examining the effects of elementary class size and composition, but previous versions of this paper contain results based on the earlier tests. These results are available from the author.

and twelve years of district-level data (from 1986–1987 to 1997–1998). I mainly show results for the fourth and sixth grade tests since they are closely linked to elementary class size, but similar eighth grade results are shown in Hoxby [1998] and are available from the author. In most years, class size is reported by multiple sources, and cross-checks of those sources suggest that it is accurate.[20] Average class size in Connecticut is about 21 students, and its standard deviation is about 5.5 students, but class size ranges widely. The first percentile is 8 students, and the ninety-ninth percentile is 34 students. While Connecticut is not unique in having appropriate data, few other states have similarly propitious conditions and long panels of the relevant data.

Table I shows the structure of the Connecticut data by cohort. Each cohort is described by its likely graduating class—for instance, one expects that children who enter sixth grade in the fall of 1991 will be in the June 1998 graduating class. Enrollment, class size, and some of the achievement data are available by school, by grade, and by cohort. The kindergarten population data and some of the achievement data are available by district, by grade, and by cohort. I have 24 years of enrollment data, so I estimate the enrollment residuals based on all 24 years of data. The large number of years allows me to get more precise estimates of the residuals.

The tests are administered at the beginning of each school year (September). Thus, the fourth grade tests may be affected by class sizes in the first through third grades, but they are unlikely to be affected by fourth grade class size. Similarly, class sizes in first through fifth grades are relevant for the sixth grade tests. Each equation has, as explanatory variables, the class sizes that could have affected the dependent variable. However, note that

[20] One might worry about error in the measure of class size, especially because measurement error can be exacerbated by first-differencing. I have verified, however, that there is little measurement error in class size by examining multiple, independent reports on class sizes. See the notes to the Appendix. More importantly, a cohort usually experiences small or large class size for several years running, so almost none of the difference between the average class size experienced by one cohort and the next (within a school) is likely to be caused by measurement error. If there were measurement error, it would average out over a few grades for a cohort. Finally, average class size is instrumented by predicted average class size, and this should remedy measurement error bias.

Measurement error in the dependent variable would show up in the standard errors, which are very small. The dependent variables are measured with error because the tests are administered in September, but fortunately, there is low student turnover. This is because families time their moves to coincide with school changeovers. Thus, a district with moderate turnover has low within-school turnover. In Connecticut in 1997–1998, the mean elementary school had 93 percent of its students return.

TABLE I
STRUCTURE OF THE DATA SET

		District-grade level data						School-grade level data				
				Tests in grade						Tests in grade		
Graduating class	Kinder. cohort	Enrollment	Class size	4	6	8	9	Enrollment	Class size	4	6	8
1983	✓	✓	✓				✓	✓				
1984	✓	✓	✓				✓	✓				
1985	✓	✓	✓				✓	✓				
1986	✓	✓	✓				✓	✓				
1987	✓	✓	✓				✓	✓				
1988	✓	✓	✓				✓	✓				
1989	✓	✓	✓					✓				
1990	✓	✓	✓					✓				
1991	✓	✓	✓			✓		✓				
1992	✓	✓	✓			✓		✓				
1993	✓	✓	✓		✓	✓		✓				
1994	✓	✓	✓		✓	✓		✓				
1995	✓	✓	✓	✓	✓	✓		✓				
1996		✓	✓	✓	✓	✓		✓	✓			
1997		✓	✓	✓	✓	✓		✓	✓			✓
1998		✓	✓	✓	✓	✓		✓	✓			✓
1999		✓	✓	✓	✓	✓		✓	✓		✓	✓
2000		✓	✓	✓	✓	✓		✓	✓		✓	✓
2001		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
2002		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
2003		✓	✓	✓	✓			✓	✓	✓	✓	
2004		✓	✓	✓	✓			✓	✓	✓	✓	
2005		✓	✓	✓				✓	✓	✓		
2006		✓	✓	✓				✓	✓	✓		

A cohort's "graduating class" is the calendar year in which it would be expected to obtain its high school diploma if its members graduated on time. For instance, if a student obtains his high school diploma in June 1998, then his graduating class is 1998. The school-level panel is slightly unbalanced because grades are occasionally moved between schools within a district. There are 3504 school-level observations of first and second grades (6 years times approximately 584 schools); 3464 school-level observations of third grades (6 years times approximately 577 schools); 3404 school-level of observations of fourth grades (6 years times approximately 567 schools); 3071 school-level observations of fifth grades (6 years times approximately 511 schools); and 1150 school-level observations of sixth grades (6 years times approximately 192 schools). Connecticut has 146 elementary districts, and there are 1752 observations in district-level regressions (12 years times 146 districts).

most cohorts experience similar class sizes in the first through sixth grades. Unusually large cohorts tend consistently to experience large class sizes, and unusually small cohorts tend consistently to experience small class sizes.[21]

[21] Statistical evidence for the last statement can be obtained by examining the correlation between, say, a cohort's first grade enrollment residual and its fifth

Every school district in Connecticut was surveyed about its maximum and minimum class size rules, teachers' aides, and mixed-grade classes. A copy of the survey is available from the author. Responses were gathered by mail, e-mail, telephone, and fax, and the researchers spoke to multiple people in most districts, although the most common respondent (by far) was the district superintendent and the second most common was a representative of the school board. The key features of the responses are as follows. Information was obtained from every district, and superintendents were queried about rules in their district over the past decade. Both maximum and minimum class size rules varied among the districts, but the modal maximum class size was 25 and the modal minimum class size was 15. Only five districts reported a change in their rules, and the changes were very modest (from maximum class size of 27 to 25, for instance).[22] The lack of changes was explained by a number of superintendents, who reported that their districts' rules had been set during the early 1980s when student populations in Connecticut were at their nadir. During subsequent years, when student populations began to grow rapidly, most districts found that *maintaining* their rules was sufficiently challenging. About one-third of districts claimed that they did not have a minimum class size rule because such a rule would never bind. Empirically, it turned out to be true that districts that claimed that they did not need the rule were districts that had steady increases in their school-aged population for the entire period. In such districts, minimum class size rules never need to be used.[23]

Districts' answers to the questions about teachers' aides and mixed-grade classes were relatively uniform. Although teachers' aides and mixed-grade classes are sometimes used for pedagogical purposes (aides are used especially for special education), they are rarely used as a method of managing too-large classes.[24]

grade enrollment residual. (Any pair of grades is suitable.) The enrollment residuals are computed using grade-school-specific regressions based on equation (3) with a quartic in time. The enrollment residuals for any pair of grades are computed independently. Nevertheless, there is a correlation of about 0.85 between pairs of residuals for a cohort.

[22] I do not use these changes in rules. For the five districts in question, I effectively divide each district into two: a "before change" district and an "after change" district.

[23] Two districts stated that they did not have any maximum class rules because they would never need to be applied. These statements were confirmed by the empirical evidence: the two districts in question have small cohort sizes for each grade (almost always under 20).

[24] Most districts were vehemently opposed to the use of aides or mixed-grade classes as a regular remedy for too-large or too-small classes.

The raw scores for Connecticut's tests are not intuitive, so I form dependent variables for the regressions by dividing each test score by the standard deviation of schools' scores on that test in Connecticut.[25] For the purpose of interpretation, it is convenient that a standard deviation on each test corresponds roughly to the state's idea of a mastery level. For instance, on the math test the difference between being "at the state's goal" and "slightly below the state's goal" is little more than one standard deviation. Similarly, the difference between being "slightly below the state's goal" and "below the state's goal" is about one standard deviation, and the difference between being "below the state's goal" and "well below the state's goal" is about one standard deviation.

All the data used are publicly available and were obtained from the Connecticut Department of Education or its publications. The Appendix shows unweighted descriptive statistics of the data set, where an observation is a school.

V. Some Illustrative Graphs

Graphs for individual schools can provide intuition about the empirical strategy and the results. I consider three schools in Connecticut, chosen for their illustrative value rather than their representativeness. School A has one classroom per grade; school B has either one or two classrooms per grade, depending on enrollment; and school C has either two or three classrooms per grade, depending on enrollment.[26] Each of Figures I through III shows a school's enrollment and class size in the fifth grade, by cohort. I selected the fifth grade because students are tested at the beginning of the sixth grade year, but it would not have mattered much if I had selected another grade. Figure I shows that, in school A, enrollment and class size were identically equal for

[25] The standard deviations in schools' scores that I use come from technical reports written by the test makers [Harcourt-Brace Educational Measurement] and distributed by the state's Board of Education. The standard deviation in students' scores on a test are about 25 percent greater than the standard deviations in schools' scores. If I were to use the standard deviation in students' scores, the estimates would appear to be even more precise. For each test, I used the median standard deviation among the years for which I have test data. The standard deviation on a test does not differ much from year to year, however, and the results are not sensitive to dividing each test score by the standard deviation for its year. The technical reports are a good guide to the scoring of the tests, which was changed once ("first generation" versus "second generation" in the terminology of tests). In the regressions I do not use data across years in which the scoring on a test changed, and the year effects in the regressions pick up idiosyncratic changes in the test from year to year.

[26] The names of schools A, B, and C are available from the author.

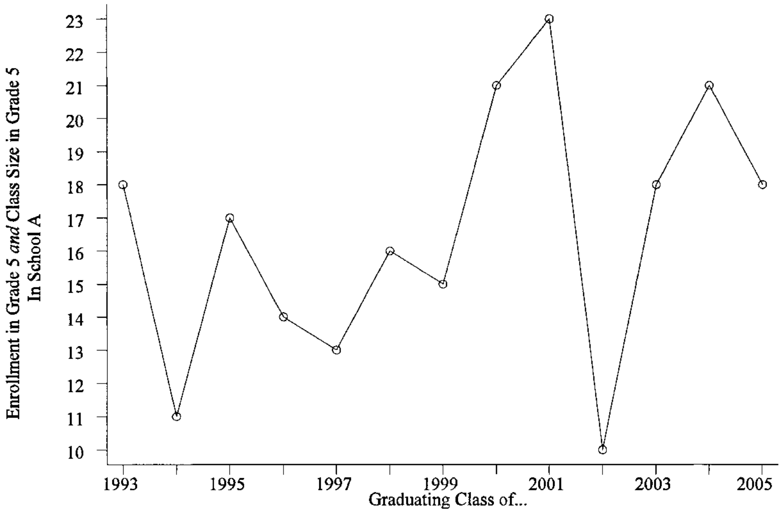

FIGURE I
Enrollment and Class Size in School A

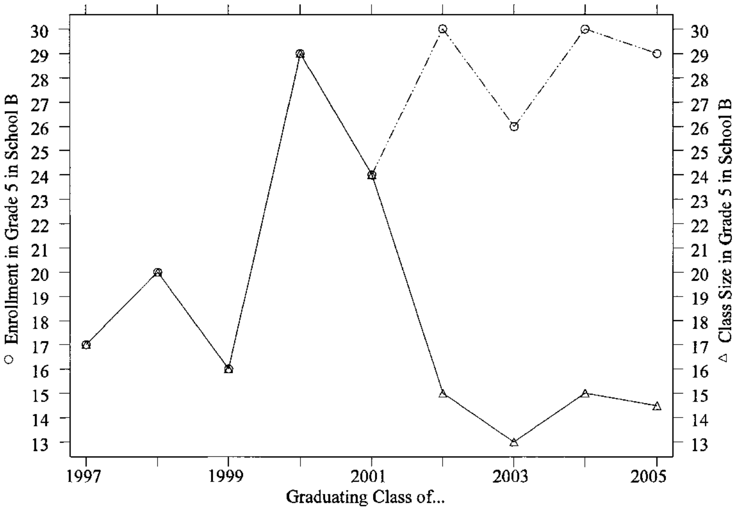

FIGURE II
Enrollment and Class Size in School B

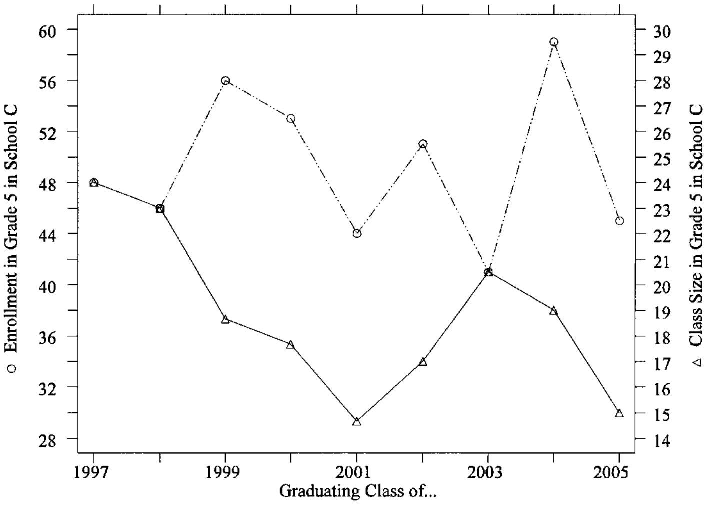

FIGURE III
Enrollment and Class Size in School C

every cohort, and class size varied between 10 and 23. Figure II shows that, in school B, enrollment and class size were identically equal for the first five cohorts, up to the graduating class of 2001. For these first five cohorts, class size varied between 16 and 29 students. The graduating class of 2002 had enrollment of 30 students, however, and school B is in a district with a maximum class size is 29. Therefore, school B added a second fifth grade classroom for the graduating class of 2002. Thereafter, even though enrollment fell back below 30 students, school B maintained two fifth grade classrooms because enrollment never fell so far that the district's minimum class size rule was triggered. The graduating classes of 2002–2005 experienced class sizes ranging from thirteen to fifteen. Figure III shows school C, which began with two fifth grade classrooms and class size of 24. However, there were 56 students in the graduating class of 1998, and school C is a district with maximum class size of 25. Therefore, school C added a fifth grade classroom for that cohort and kept the third classroom until the graduating class of 2003, which had enrollment of only 40 students. The minimum class size in school C's district is fourteen, so the rule was triggered, and school C went back to having only two fifth grade classrooms. The next cohort, however, had 59 students, and the third fifth grade classroom was reinstated. Overall, class size varied from 14 to 24 students in school C.

All three figures illustrate the variation that is useful for the first identification method, which uses the variation in enrollment that does not appear to be part of a trend and that does not trigger a change in the number of classrooms. School A is a nice, simple example because, although it appears to have an upward trend in enrollment, it is obvious that much of the year-to-year variation in enrollment is not systematic. Figures II and III illustrate the variation that is useful for the second identification method, which uses the changes in class size that occur when within-school variation in enrollment triggers a maximum or minimum class size rule. All three figures show that class size varies over a range that covers the policy range very fully. Just in these three schools, class size varies from 10 to 29.

Figures IV through VI superimpose each school's average sixth grade reading scores on its fifth grade class size. If reducing class size improved reading scores, then we would expect to see the two lines generally move in opposite directions, like mirror images of one another. But, it is difficult to discern any pattern

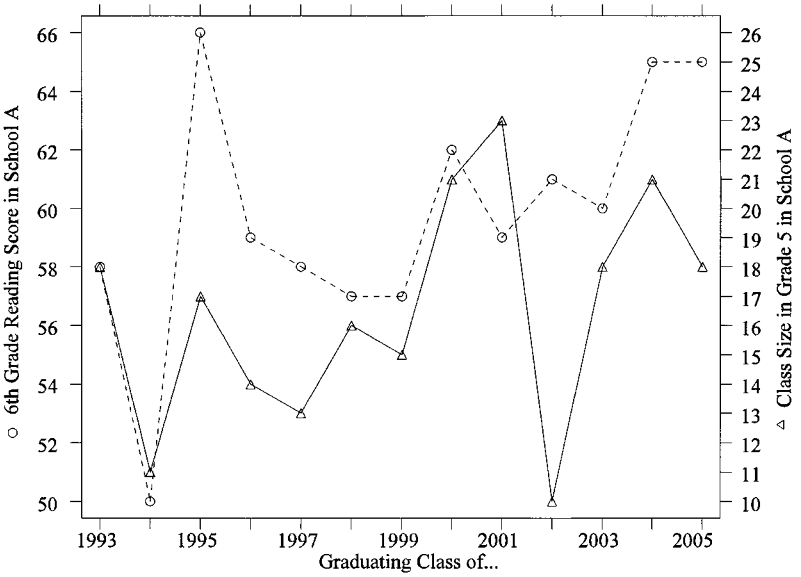

FIGURE IV
Class Size and Reading Scores in School A

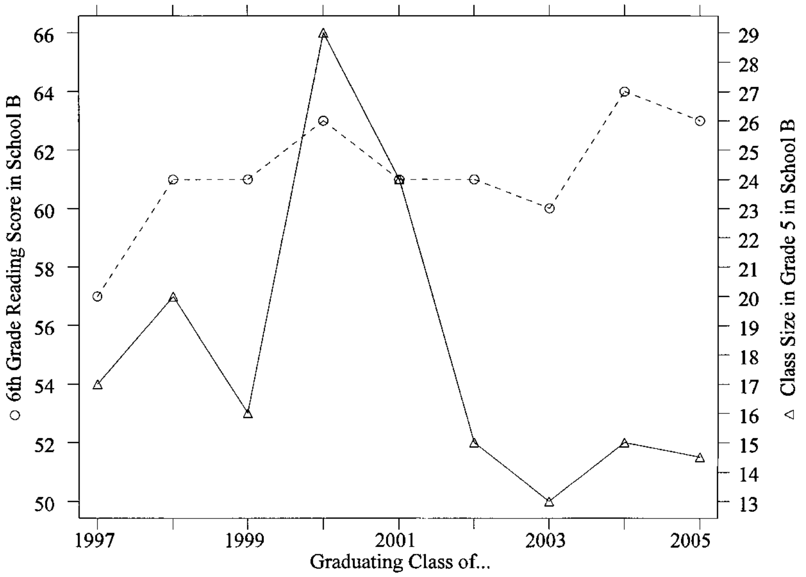

FIGURE V
Class Size and Reading Scores in School B

linking reading scores and class size. The same can be said for math scores and writing scores, which are not shown here. Looking at these three schools, however, is hardly a systematic way of determining whether there is a significant relationship between achievement and class size. There is a need for regression analysis.

VI. RESULTS

In this section I examine the effects of class size on achievement. Before showing the results for the two identification methods described above, I show results for a few methods that are commonly used despite having the identification problems described in Section II. These results give one a sense of what the data would show if one were to apply typical methods naively.

1. Results from Commonly Used Methods of Identification

Each cell in Table II shows the estimated coefficient on class size from a separate regression. The columns define the specification of the regression, and the rows show results for different

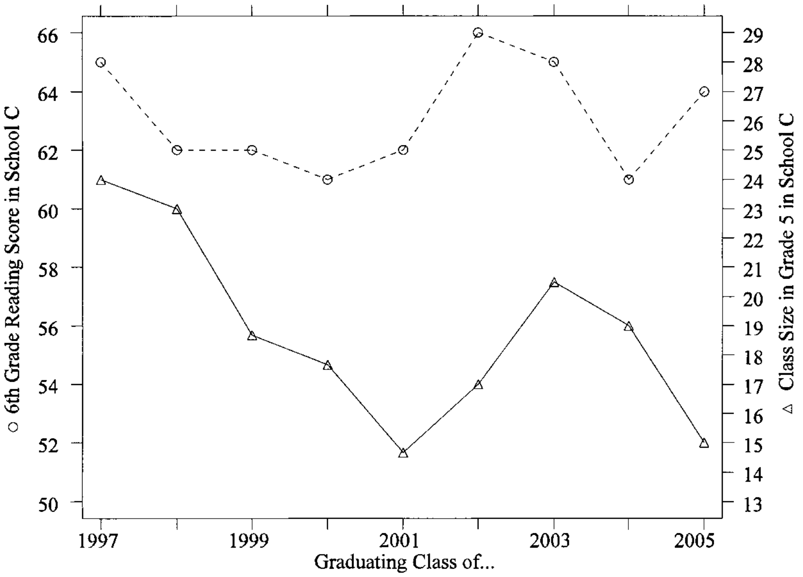

FIGURE VI
Class Size and Reading Scores in School C

dependent variables. For instance, the number in the upper-left-hand cell is the effect of log average class size in grades 1 through 3 on fourth grade math scores using a specification that pools observations across schools and cohorts (with cohort fixed effects). This naive specification is likely to produce estimates that are biased by correlation between class size and unobserved parent and community attributes. Parents with unobserved good characteristics are likely to choose schools with small class sizes and communities with unobserved good qualities. In fact, the estimates in the first six rows of column I are all negative and highly statistically significant. (I discuss the bottom three rows below.) If one were to give the estimates credence, one would say that the coefficient in the first cell indicates that a 10 percent reduction in class size in grades 1 through 3 improves fourth grade math scores by 0.1468 (about 15 percent) of a standard deviation. Other coefficients in column I are similar: a 10 percent reduction in class size in grades 1 through 5 appears to improve sixth grade math scores by about 13 percent of a standard deviation.

In column II, I augment the equation by adding district-level demographic variables from the 1990 census: median household

TABLE II
NAIVE ESTIMATES OF THE EFFECTS OF CLASS SIZE ON STUDENT TEST SCORES
Each cell contains the estimate from a separate regression (and its standard error in parentheses).

Dependent variable	Independent variable	I Cohort fixed effects	II Cohort fixed effects & demographic controls
fourth grade math score	log avg class size through grade 3	−1.4675 (0.2067)	−0.1028 (0.0994)
fourth grade reading score	log avg class size through grade 3	−1.1532 (0.1450)	−0.1338 (0.0752)
fourth grade writing score	log avg class size through grade 3	−0.5872 (0.0919)	−0.0301 (0.0578)
sixth grade math score	log avg class size through grade 5	−1.3141 (0.2788)	−0.1364 (0.1209)
sixth grade reading score	log avg class size through grade 5	−1.4043 (0.2771)	−0.1821 (0.1162)
sixth grade writing score	log avg class size through grade 5	−0.5571 (0.1409)	−0.0497 (0.0907)
sixth—fourth grade math score	avg class size in fourth and fifth grds	0.1081 (0.0829)	−0.1335 (0.0722)
sixth—fourth grade reading score	avg class size in fourth and fifth grds	−0.2645 (0.0581)	−0.1572 (0.0498)
sixth—fourth grade writing score	avg class size in fourth and fifth grds	−0.1980 (0.0848)	−0.2950 (0.0968)

Source is author's calculations based on Connecticut data set. The regressions are OLS, are weighted by number of students over whom the dependent variable is averaged, and include a fixed effect for each cohort. Standard errors are in parentheses and adjusted for the grouped nature of the data (multiple observations on each school). The number of observations in the regressions for grades 1 through 6 is, respectively, 3504, 3504, 3464, 3404, 3071, 1150 (see the notes to Table I). The dependent variables are formed by dividing the average test score by the standard deviation of Connecticut students' scores on that test. Thus, the coefficients show how test scores, *measured in standard deviations*, change with the log of class size. The demographic controls in column II are median household income, percentage of the population in poverty, percentage of adults who are high school graduates, percentage of adults who are college graduates, percentage of the population who are African-American, and percentage of the population who are Hispanic.

income, the percentage of the population in poverty, the percentage of adults who are high school and college graduates, and the percentages of the population who are African-American and Hispanic. (These variables are observed at the district level only in decennial census years.) These controls for observed parent and community characteristics greatly attenuate the estimated effect of class size on test scores, but the estimates are still all negative in sign, and two of the six estimates in the first six rows are statistically significant at the 10 percent level. In fact, the estimates shown in column II are similar to many of the estimates that have generated empirical controversy. They are of mixed or marginal statistical significance, and the effects are *small*. The equations control for some observed demographics, but it is not clear that the remaining variation in class size comes from exogenous sources. At least some of the remaining variation is likely to be due to unobserved demographics that are correlated with class size in much the same way as the observed demographics are correlated with class size: parents with demographics that are beneficial for achievement choose districts with smaller class sizes, producing results biased toward finding that class size reductions are efficacious. On the other hand, other demographic controls being equal, the schools with lower class size may be those that are receiving compensatory funds to reduce class size *because* their students have unusually low achievement. This would produce results biased against finding that class size reductions are efficacious.

The bottom three rows of Table II show what is usually called a value-added specification. The difference between a cohort's sixth grade and fourth grade test scores is regressed on the log of the average class sizes that they experience between the two tests: fourth and fifth grade. Such specifications are often thought to control for all the effects of family background and neighborhood, through the earlier test score. It is far from obvious, however, that such claims are valid. Unobserved background may affect the *growth* of a student's achievement; unobserved background need not be fully incorporated by the *level* of a student's prior achievement. In other words, parents who provide a lot of learning resources at home are likely to help their children learn more in every grade, for every bundle of resources that the child gets at school. The same parents are likely to put their children in schools with small class size.[27] In short, value-added estimates may be biased either negatively or positively, but it is likely that the preponderance of the bias favors class size appearing to be efficacious. In fact, five of the six estimates in the bottom three rows of Table II are negative and statistically significantly different from zero at the 5 percent level. If one were to give the

[27] The problem would not be alleviated if I regressed the change in test scores on the *change* in the class size (average class size in fourth and fifth grades minus average class size in first through third grades). Most of the changes in class size that a cohort experiences would *not* be random. It would be the result of reactions to the cohort's own achievement or the result of systematic changes in the school's environment.

estimates credence, one would say that a 10 percent reduction in class size in grades 4 and 5 makes reading scores rise (between the fourth and sixth grades) by about 16 percent of a standard deviation, controlling for observable demographics.

The fundamental problem with all of the specifications in Table II is that they eliminate one source of suspect variation, only to have more obscure sources become dominant. When considering a policy variable like class size, where the vast majority of the variation comes from suspect sources, it is more effective to start with sources of variation that are known to be exogenous and work from there. This is the logic behind the two identification methods advanced in this paper.

2. Results from the First Identification Method

The first identification method attempts to use random variation in the school-aged population, and the strategy is implemented by instrumenting for class size with enrollment residuals or kindergarten cohort residuals. Table III shows coefficient estimates from the first-stage equation (equation (7)). Each cell represents a separate regression, and each contains the effect of log (u) on the log of class size. Each column heading describes the specification, and each row label describes the grade level for which class size is being estimated. For instance, the upper-left-hand cell contains the estimated coefficient on log (u) from a regression that is based on school-level observations of first grade class size and school-level observations of first grade enrollment residuals, where the enrollment residuals are calculated using a version of equation (3) that has an intercept and a *linear* time trend. The estimated coefficients in the first column range from 0.8566 to 0.9773. They suggest that a random 10 percent increase in enrollment raises class size by between 8.6 and 9.7 percent—in other words, a little less than 1-for-1. The probable reason why the coefficients are not even closer to 1 is that the enrollment residuals are measured with error (they are just estimates, after all). There is, thus, a little attenuation bias. Column II contains the estimated coefficients on log (u) from regressions that are based on school-level observations of class size and enrollment residuals, where the enrollment residuals are calculated using a version of equation (3) that has an intercept and a *quartic* time trend. The estimates shown in column II are very similar to those shown in column I: they range from 0.8546 to 0.9799. The estimates in column III, which is estimated at the district level,

TABLE III

COEFFICIENTS FROM FIRST-STAGE REGRESSIONS FOR IDENTIFICATION METHOD 1

Each cell contains the estimate from a separate regression (and its standard error in parentheses).

	I	II	IV	V
	Explanatory variable			
Dependent variable	Residual log enrollment (school-level, linear time trend removed)	Residual log enrollment (school-level, quartic time trend removed)	Residual log enrollment (district-level, quartic time trend removed)	Residual log kindergarten cohort (district-level, quartic time trend removed)
log first grade class size	0.8566	0.8546	0.7620	0.6186
	(0.0110)	(0.0210)	(0.0349)	(0.0452)
log second grade class size	0.7294	0.7275	0.6679	0.6416
	(0.0129)	(0.0183)	(0.0346)	(0.0423)
log third grade class size	0.8937	0.8717	0.7854	0.4557
	(0.0105)	(0.0164)	(0.0360)	(0.0459)
log fourth grade class size	0.9419	0.8920	0.7887	0.3786
	(0.0098)	(0.0309)	(0.0365)	(0.0480)
log fifth grade class size	0.9039	0.8678	0.7027	0.3953
	(0.0128)	(0.0197)	(0.0345)	(0.0434)
log sixth grade class size	0.9773	0.8669	0.8356	0.3099
	(0.0111)	(0.0290)	(0.0499)	(0.0623)
fixed effects for each "school · expected number of classes in the grade" combination	yes	yes		
fixed effects for each "district · expected number of classes in the grade" combination			yes	yes

Source is author's calculations using Connecticut data set. Identification Method 1 attempts to use random variation, over time, in the population of students who belong to a grade in a school as an instrument for class size in that grade in that school. Each first-stage regression has, as its dependent variable, the log of class size in a grade. Each regression has, as its key explanatory variable, an estimate of the part of the grade's population that is due to random variation. For instance, each explanatory variable in column I is the residual from a regression of enrollment in a grade in a school on a constant and a linear time trend. The residuals come from *separate* regressions for each grade in each school. Each first-stage regression contains a fixed effect for each school-expected number of classes combination. These fixed effects ensure that the monotonicity condition for instrumental variables is fulfilled. See Section III for further explanation. In the school level regressions, the number of observations is 3404 in fourth grade regressions, and 1150 in sixth grade regressions. In the district level regressions there are 1752 observations (see the notes to Table I). If the independent variable is class size in the most recent grade, instead of average class size in grades that precede the test, then the results for the specification in column II are −0.1304 (0.0980), −0.1204 (0.0747), 0.1550 (0.0901), 0.0304 (0.1167), −0.0330 (0.1084), and 0.0925 (0.1537).

are also similar: they range from 0.7027 to 0.8356. There is more attenuation bias in the district-level regressions because the district-level enrollment residuals are a less precise measure of the random fluctuations in enrollment experienced by any given school in the district. Finally, column IV of Table II shows results based on district-level kindergarten cohort residuals. As one expects, the coefficients are somewhat lower than those of column III because some children in the kindergarten cohort go to private school. Moreover, a 10 percent increase in kindergarten cohort size produces the biggest increase in first grade class size, a smaller increase in second grade class size, and so on down to sixth grade class size. One expects this because mobility into and out of the district makes kindergarten-cohort size more important for the earlier elementary grades.

Overall, the first-stage regressions suggest that enrollment residuals are strong instruments for class size: the *t*-statistics in the first two columns are all greater than 40. District-level and kindergarten-cohort residuals are less strong as instruments, but still strong enough: the *t*-statistics are generally much greater than 10. Also, the coefficients accord with expectations, which suggests that the residuals are being estimated in a reasonable fashion. Residuals from school-specific intercepts and *quartic* time trends are my preferred set of residuals. Moving from a quartic polynomial to a high-order polynomials adds a negligible amount of explanatory power and produces residuals that are not discernibly different.

I use predicted class size based on the equations shown in Table III to form independent variables for the second-stage equations. For instance, to form a prediction of the log average class size that a cohort in a school experiences in grades 1 through 3, I compute the average of that cohort's predicted class size in grades 1, 2, and 3, and I take the log of the result. I compute the log of the average predicted class size, not the average of the log predicted class sizes. I take account of "feeder schools"—for instance, student from two first-to-fourth grade schools may attend the same fifth-sixth grade school.

Table IV contains the main class size results for the first identification method. Each cell contains an estimate from a separate regression. Column I uses first-stage regressions in column I of Table III, column II uses first-stage regressions in column II of Table III, and so on.

Before considering the estimated coefficients, note that the

TABLE IV
BASIC RESULTS FROM IDENTIFICATION METHOD 1: 2SLS ESTIMATES OF THE EFFECTS OF CLASS SIZE ON STUDENT TEST SCORES
Each cell contains the estimate from a separate regression (with its standard error in parentheses).

		I	II	III	IV
		Class size is predicted using first stage regressions from			
Dependent variable	Independent variable is the prediction of:	Column I of Table III	Column II of Table III	Column III of Table III	Column IV of Table III
fourth grade math score	log avg class size through grade 3	0.0664 (0.1069)	−0.0845 (0.1227)	0.1245 (0.2100)	0.2203 (0.1537)
fourth grade reading score	log avg class size through grade 3	−0.0736 (0.0759)	−0.1027 (0.0870)	−0.1513 (0.1643)	0.1260 (0.1084)
fourth grade writing score	log avg class size through grade 3	0.1364 (0.1085)	0.1871 (0.1214)	−0.0198 (0.1472)	0.0332 (0.1061)
sixth grade math score	log avg class size through grade 5	0.0496 (0.1367)	0.0394 (0.1578)	−0.0522 (0.1346)	0.2059 (0.1714)
sixth grade reading score	log avg class size through grade 5	−0.0174 (0.1247)	0.1288 (0.1462)	−0.0152 (0.1063)	0.0843 (0.1410)
sixth grade writing score	log avg class size through grade 5	0.0675 (0.1769)	0.0494 (0.2077)	−0.1384 (0.1166)	−0.1003 (0.1562)
fixed effects for each “school · expected number of classes in the grade” group		yes	yes		
fixed effects for each “district · expected number of classes in the grade” group				yes	yes
fixed effects for each cohort		yes	yes	yes	yes

Standard errors are correct for 2SLS. The regressions are weighted by the number of students over whom the dependent variable is averaged. In the school level regressions the number of observations is 3404 in fourth grade regressions, and 1150 in sixth grade regressions. In the district level regressions there are 1752 observations (see the notes to Table I). Predicted class size is computed using the first-stage regressions shown in Table III. The dependent variables are formed by dividing the average test score by the standard deviation of students' scores on that test in Connecticut. Thus, the coefficients show how test scores, *measured in standard deviations*, change with the log of class size.

standard errors are small. In the school-level regressions (columns I and II) the standard errors are so small that if a 10 percent reduction in class size were to change test scores by just 2 to 4 percent of a standard deviation, the change would be statistically significant at the 5 percent level. In the district-level regressions (columns III and IV) the standard errors are slightly higher, but if a 10 percent reduction in class size were to change test scores by just 3 to 4 percent of a standard deviation, the change would be statistically significant at the 5 percent level. In other words, if reducing class size by 10 percent made students move just 2 to 4 percent closer to mastering the state's next level of proficiency, the improvement would be statistically significant. The random variation in class size has considerable power to identify achievement gains.

Despite this propitious situation, the estimates in columns I through IV do not show that smaller class sizes produce achievement gains. The estimates are mixed in sign, and none is statistically significant at the 5 percent level. One would not wish for smaller standard errors because as many results with the "wrong" as with the "right" sign would become statistically significant. The simplest interpretation of Table IV is straightforward: given the standard errors, the effect of reducing class size is rather precisely estimated to be close to zero. Because all of the estimates are close to zero, the four specifications do not seem very different, but in fact we should remember that they use different enrollment residuals as the estimates as log (u). In particular, the column III estimates do not allow parents' moving students within the district to produce bias, and the column IV estimates do not allow parents' moving students to other districts or to private schools to produce bias.

Given that Table IV presents "well-estimated zeros," one is naturally drawn to estimate a variety of alternative specifications to see if and when class size matters. I can show only a fraction of the specifications I estimated. The notes to Table IV present results for class size in the most recent grade. In Hoxby [1998] I explore numerous other specifications and demonstrate that results much like those in Table IV are obtained if the independent variable is class size in grade 1, class size in grades 1 and 2, class size with a spline (with a break at class size of 23), an indicator for students' ever having experienced class size below 15, or an indicator for students' ever having experienced class size above 30. In Table V, I show two alternative specifications that are especially likely to be interesting, given the empirical literature on class size. Krueger [1999], Hanushek, Kain, and Rivkin [1998], and Ferguson [1998] argue that the reductions in class size are more efficacious in schools that serve students who are low-income or minorities. African-Americans are the most important minority group in Connecticut, so I examine results that differ by whether the school's students come from low, medium, or high

TABLE V
ADDITIONAL RESULTS FROM IDENTIFICATION METHOD 1: 2SLS ESTIMATES USING SPLINE SPECIFICATIONS
Each cell contains the estimate from a separate regression (with its standard error in parentheses).

		Spline: each row is a regression			Spline: each row is a regression		
Dependent variable[2]	Independent variable	Low inc	Med inc	High inc	High %blk	Med %blk	Low %blk
fourth grade math score	log avg class size through grade 3	−0.0455	−0.0877	−0.1490	−0.1837	−0.1250	0.1618
		(0.1425)	(0.1299)	(0.1409)	(0.1455)	(0.1445)	(0.2863)
fourth grade reading score	log avg class size through grade 3	−0.1218	−0.1638	−0.2052*	−0.1500	−0.0855	−0.3557
		(0.0997)	(0.1009)	(0.0986)	(0.1016)	(0.1009)	(0.2000)
fourth grade writing score	log avg class size through grade 3	0.2213	0.2002	0.1323	0.2253	0.2002	0.0899
		(0.1386)	(0.1277)	(0.1451)	(0.1423)	(0.1419)	(0.2931)
sixth grade math score	log avg class size through grade 5	0.3121	−0.0072	−0.0203	0.0943	0.1478	0.1501
		(0.2627)	(0.2635)	(0.4847)	(0.3035)	(0.2523)	(0.3945)
sixth grade reading score	log avg class size through grade 5	0.2811	0.0840	0.0028	0.5378	0.1493	0.2619
		(0.2416)	(0.2437)	(0.4487)	(0.3651)	(0.2324)	(0.2810)
sixth grade writing score	log avg class size through grade 5	0.3074	0.0482	−1.0847	0.4680	−0.2835	0.0160
		(0.3320)	(0.3350)	(0.6166)	(0.3865)	(0.3197)	(0.5022)

See notes to Table IV except for the following details specific to Table V. The regressions in Table V are the same as the regression reported in column II of Table IV, except that they have splines created by interacting the class size variable with indicator variables for a characteristic of the school district. The indicator for a low income district is equal to 1 if the district has median household income less than or equal to the twenty-fifth percentile of median household income in Connecticut districts; 0 otherwise. The indicator for a medium income district is equal to 1 if the district has median household income greater than the twenty-fifth percentile and less than seventy-fifth percentile; 0 otherwise. The indicator for a high income district is equal to 1 if the district has median household income greater than or equal to the seventy-fifth percentile; 0 otherwise. The indicators for low, medium, and high percent African-American are constructed similarly around the twenty-fifth and seventy-fifth percentiles of the percentage of the population that is African-American in Connecticut districts.

income families and by whether the school's share of students who are African-American is high, medium, or low.[28]

Specifically, in Table V, I estimate the specification from column II of Table IV, except that I allow class size to have different coefficients for schools that fit into different groups. I first divide schools into groups where the "low income" group is districts with per capital income at or below the twenty-fifth percentile of per capita income in Connecticut (15,454 dollars in 1990), the "medium income" group has per capita income above the twenty-fifth percentile and below the seventy-fifth percentile (23,075 dollars in 1990), and the "high income" group has per capita income at or above the seventy-fifth percentile. These divisions produce the estimates in columns I through III. With one exception, none of the coefficients is statistically significantly different from zero, although the standard errors are still small enough to generally identify improvements in achievement as small as 3 to 6 percent of a standard deviation for a 10 percent reduction in class size. In no case is the point estimate for high income schools statistically significantly different from the point estimate for low income schools. Also, the discernible pattern of the point estimates does not suggest that class size reductions are more efficacious in schools that serve low income students (if anything, the pattern suggests the opposite). The only statistically significant estimate suggests that class size reductions improve fourth grade reading scores in schools that serve students from high income backgrounds. Perhaps teachers who work in such schools are more likely to make good use of class size reductions, or high income parents are more likely to ensure that their "slow reader" gets individual attention when class size is small.

I next divide schools into groups where the "high percentage African-American" group contains districts with percent African-American at or above the seventy-fifth percentile of districts' percent black in Connecticut (17 percent in 1990), the "medium" group contains districts with percent African-American below the seventy-fifth percentile and above the twenty-fifth percentile (1 percent in 1990), and the "low" group contains districts with percent African-American at or below the twenty-fifth percentile. In Connecticut, African-American households are concentrated in

[28] In 1990, 11 percent of Connecticut's school-aged population was African-American, slightly less than 9 percent was Hispanic, and slightly less than 2 percent was Asian.

urban districts, so the "high" group may also be thought of as the urban group. These divisions produce the estimates in columns IV through VI. None of the coefficients is statistically significantly different from zero, although the standard errors are still small enough to generally identify improvements in achievement as small as 3 to 6 percent of a standard deviation for a 10 percent reduction in class size. In no case is the estimate for high percent African-American schools statistically significantly different from the estimate for low African-American schools, and there is no discernible pattern in the point estimates.

In summary, the estimates in Tables IV and V suggest that class size reductions are not efficacious for improving student achievement. The estimates do not confirm the hypothesis that class size reductions are more efficacious in districts that contain low income or African-American students.

3. Results from the Cross-Section Regression Discontinuity Method

Now consider changes in class size that occur when a school changes the number of classes in a grade. In Table VI, I treat the data as though they were cross-section data, estimate the predicted class size function for each school based on its district's maximum class size and equation (9), and use the log of predicted class size as an instrument for log class size. (This method does not lend itself to examining the effects of class size in multiple grades, so I use class size in the grade immediately prior to the test.) Recall that the cross-section approach is likely to produce unbiased results only when the sample is narrowed to the observations just on either side of a maximum class size threshold. In column I, I use the entirety of the predicted class size function and expect to produce biased results, since most of the function reflects permanent characteristics of the school. In column II, I use only the observations that are within four students of a discontinuity, so the results should be less biased. In column III, I use only the observations that are *at* a discontinuity, and I expect the results to be unbiased.

Consider first the number of observations in each regression, shown at the bottom of Table VI. As one narrows in on the discontinuities, the number of observations in the fourth grade regressions falls from 1953 in column I to 76 in column III. The number of observations in the sixth grade regressions falls from 1011 in column I to 37 in column III. As the number of observa-

TABLE VI
IV ESTIMATES OF THE EFFECT OF CLASS SIZE, GENERATED BY *CROSS-SECTION* REGRESSION DISCONTINUITY
Each cell contains the estimate from a separate regression.

	I	II	III
	The predicted class size function is used:		
Dependent variable	In its entirety	Within 4 students of a discontinuity	Solely at the discontinuities
fourth grade math score	−0.0503	−0.0972	−0.0506
	(0.0229)	(0.0593)	(0.1060)
fourth grade reading score	−0.0423	−0.0856	−0.0746
	(0.0166)	(0.0454)	(0.0821)
fourth grade writing score	−0.0137	−0.0082	−0.0211
	(0.0130)	(0.0321)	(0.0372)
sixth grade math score	−0.0922	−0.0992	0.0674
	(0.0496)	(0.0872)	(0.1220)
sixth grade reading score	−0.1042	−0.1496	0.0250
	(0.0511)	(0.0992)	(0.0796)
sixth grade writing score	−0.0301	−0.0181	0.0159
	(0.0241)	(0.0401)	(0.0498)
number of observations in fourth grade regressions	1953	703	76
number of observations in sixth grade regressions	1011	374	36

The source is author's calculations based on the Connecticut data set. The regressions are weighted by the typical number of observations over which the dependent variable is averaged. The dependent variables are formed by dividing the average test score by the overall standard deviation of scores on that test in Connecticut. The independent variable is class size in most recent grade, instrumented by predicted class size. The equation contains a fixed effect for each cohort. The cross-section method treats the Connecticut data as though they were cross-section data and actual changes in the number of classes were not observed. The predicted class size function uses each district's maximum class size and the formula given by equation (9). See text for further explanation.

tions falls, the standard errors rise. The standard errors in column I are such that a 10 percent reduction in class size would have to produce an improvement of 6 to 16 percent of a standard deviation for the improvement to be statistically significant. The standard errors in column III are such that a 10 percent reduction in class size would generally have to produce an improvement of 30 to 50 percent of a standard deviation for the improvement to be statistically significant. The falling number of observations and the rising standard errors demonstrate the extraordinary demands that the cross-section method puts on data when it is applied so as to ensure unbiased results.

The regressions in column I suggest that reductions in class size improve achievement significantly. Four out of the six coefficients are statistically significant at the 5 percent level, and all six coefficients have the “right” sign. If we were to interpret these results naively, we would conclude that a 10 percent reduction in third grade class size raises fourth grade math scores by about 12 percent of a standard deviation. As we narrow in on the discontinuities, however, such results disappear. In column III, where only the discontinuities are used, none of the results is close to being statistically significant, and four out of the six estimates have the “wrong” sign. Therefore, the statistically significant results in column I are generated *not* by the discontinuities in the predicted class size function, but by the suspect parts of the function.

4. Results from the Second Identification Method (the Within-School Regression Discontinuity Method)

Table VI shows results from the second identification method: the within-school regression discontinuity method. For this method I focus on events where the number of classes changed because a modest change in enrollment (smaller than 20 percent) triggered a maximum or minimum class size rule. I estimate equation (11), a first-differenced version of the achievement equation, using just the cohorts immediately before and after each event. This method is quite powerful despite the fact that it relies purely on discontinuous changes in class size driven by changes in the number of classes. Its power derives from the fact that it compares adjacent cohorts in the same school, who have little reason to be different apart from their different class size experiences. In fact, the second identification method produces standard errors so small that if a 10 percent reduction in class size were to change test scores by just 2 to 4 percent of a standard deviation, the change would be statistically significant at the 5 percent level.

Column I of Table VII includes all the events in which the number of classes changed (and affected class size) in the grade before the test. Column II includes only the events in which the number of classes changed in the same way (and affected class size) in the three grades immediately before the test. In other words, column II uses the fact that a cohort that was big enough to have a third grade class added when it entered third grade often had a class added in second grade and first grade as well. Despite the small standard errors, none of the estimates in Table VII is statistically significantly different from zero at the 5 percent level.

TABLE VII

ESTIMATES OF THE EFFECT OF CLASS SIZE, IDENTIFICATION METHOD 2 (WITHIN-SCHOOL REGRESSION DISCONTINUITY)

Each cell contains the estimate from a separate regression (with its standard error in parentheses).

	I	II
Dependent variable	Independent variable is change in class size (due to the addition or subtraction of a class) in the grade previous to the test, for the 2 adjacent cohorts	Independent variable is change in class size (due to the addition or subtraction of classes) in the 3 grades previous to the test, for the 2 adjacent cohorts
Change in fourth grade math score between two adjacent cohorts in the same school	0.0844 (0.1001)	−0.0714 (0.1605)
Change in fourth grade reading score between two adjacent cohorts in the same school	0.0468 (0.0636)	−0.0540 (0.1396)
Change in fourth grade writing score between two adjacent cohorts in the same school	0.1731 (0.0976)	0.1602 (0.1568)
Change in sixth grade math score between two adjacent cohorts in the same school	0.0126 (0.0969)	−0.0207 (0.1588)
Change in sixth grade reading score between two adjacent cohorts in the same school	−0.0468 (0.0828)	0.0238 (0.1520)
Change in sixth grade writing score between two adjacent cohorts in the same school	0.1585 (0.1300)	0.1543 (0.1901)
number of observations in fourth grade regressions	147	117
number of observations in sixth grade regressions	108	86

The source is author's calculations based on the Connecticut data set. The regressions are weighted by the typical number of observations over which the dependent variable is averaged. The dependent variables are formed by dividing the average test score by the overall standard deviation of scores on that test in Connecticut. The within-district method exploits the fact that the Connecticut data are panel data and actual changes in the number of classes within a grade within a school are observed. The equation is estimated in first-differences: the change in scores between back-to-back cohorts is regressed on the change in class size, *if* that change in class size is the result of a small change in enrollment having triggered a maximum or minimum class size rule.

One would not wish for smaller standard errors because more results with the "wrong" sign would become statistically significant than would results with the "right" sign. The best interpretation of Table VII is that the estimated effects of class size reductions are rather precisely estimated zeros.

Obviously, the 20 percent cutoff for a "small" change in enrollment is arbitrary. I have experimented with cutoffs between 35 and 15 percent, and the results are similar.[29]

VII. Interpretation

Estimates based on both identification methods indicate that class size reductions have little or no effect on achievement. The estimates are sufficiently precise that improvements that are educationally significant would be identifiable. The two identification methods are independent and thus provide checks on one another. The results are also robust to specification changes, some of which are shown above and some of which are shown in Hoxby [1998].

The estimates are based on variation in class size that occurs mainly in the range of 10 to 30 students per class. This is the relevant range for American policy, but it would be a mistake to extrapolate these results to schools in which class size is typically higher than 30. Since most schools in developing countries have class sizes higher than 30, the results in Tables IV through VII neither confirm nor contradict most developing country studies. It would also be a mistake to extrapolate these results to class sizes of less than 10. Such tiny classes are too expensive for most American districts to consider because the cost of a one-student reduction increases as class size gets smaller (cost is roughly linear in the *percentage* reduction in class size). A five-student reduction from a base of 40 raises costs by only 14.3 percent; but a five-student reduction from a base of 15 raises costs by 50 percent.

Krueger [1999] provides evidence that, in Project Star, a 10 percent reduction in class size for one year improves scores by about 10 percent of a standard deviation, a 10 percent reduction in class size for three improves scores by about 13 percent of a standard deviation (compare this with the fourth grade results in this paper), and a 10 percent reduction in class size for five years improves scores by about 17.5 percent of standard deviation

[29] These results are available from the author.

(compare this with the sixth grade results in this paper).[30] Any of these improvements would be highly statistically significant if they appeared in this paper, given this paper's standard errors.

How might one explain the contrast in results in the natural experiment and an explicit policy experiment? In both the natural experiment and policy experiment, teachers had more *opportunity* to improve achievement with smaller classes. In neither experiment did teachers receive special training to take advantage of the smaller class sizes. The difference in the results may be caused by the fact that the natural experiment varied class size but did not vary incentives, while the policy experiment varied class size and contained implicit incentives for teachers and administrators to make good use of smaller class sizes (because full enactment of the policy depended on a successful evaluation). If this is the correct interpretation of the difference in the results, then the implication is that class size reduction policies should contain built-in evaluation and incentives.

Since Connecticut school staff were unaware of the natural experiment, they could not have reacted to the evaluation. Explicit policy experiments may work differently because of Hawthorne effects or other reactive behavior on the part of participants.

One might attribute some of the difference in results to the necessarily transitory nature of population variation (from the teachers', not students', point of view). That is, teachers experience small class sizes repeatedly, but not every year. Teachers do not receive training to take advantage of smaller class sizes in a systematic way—in other words, they may not vary their primary classroom style much when they have the opportunities presented by a smaller class. This interpretation would suggest that reductions in class size should be combined with instruction for teachers that helps them modify their teaching techniques. This cannot, however, be the entire explanation. Even if she does not lecture differently to a smaller class, a teacher can devote more effort to each student during every teaching activity that has an individual element. Many of these activities are part of a teacher's basic repertoire: answering questions, correcting assignments, dealing with disciplinary problems, tutoring a student who is ahead or behind the class, talking to parents, and so on. Also, the

[30] The average "small" class in Project Star was 30 percent smaller than the average "regular" class. Students' scores increased by about 30 percent of a standard deviation on math and reading tests after one year.

Project Star results were achieved after only one year of smaller class size, and the teachers involved did not receive instruction about changing their primary teaching techniques.

VIII. Conclusions

In this study I use natural variation in the school-aged population to identify the effects of class size on student achievement. This approach has three benefits. First, the variation in class size that I study is credibly exogenous. It is not variation generated by parents' choices—choices that are affected by parents' incomes and parents' assessments of the attention their children need. Second, the actors in the natural experiment I examine were not aware of being evaluated or mindful of rewards being contingent upon the outcome. Real policies that reduce class size, such as the 1996 California initiative and the 1999 federal initiative, rarely include evaluation or repercussions (such as the funds being taken away if the policy has no effect). It is important that research mimic the incentives that exist under real policies. Third, natural population variation generates fluctuations in class size that are in the range relevant to current policy.

This study demonstrates how population variation can be used to consistently estimate the effect of class size on student achievement. I outline two independent methods for using population variation. The first method is based on isolating the credibly random component of the natural variation in population for a grade in a school. Random variation in the population generates exogenous variation in class size. The second method is based on exploiting the discontinuous changes in class size that occur when a small change in enrollment triggers a maximum or minimum class size rule and thereby changes the number of classes in a grade in a school. Both methods produce results that are appropriate for considering class size changes in the range of 10 to 30 students.

Using both methods, I find that reductions in class size have no effect on student achievement. The estimates are sufficiently precise that, if a 10 percent reduction in class size improved achievement by just 2 to 4 percent of a standard deviation, I would have found statistically significant effects in math, reading, and writing. I find no evidence that class size reductions are more efficacious in schools that contain high concentrations of low income students or African-American students.

The results just described are far less likely to suffer from omitted variables bias and endogeneity bias than are typical estimates that depend on variation in class size that is (directly or indirectly) generated by parents' decisions, teachers' decisions, administrators' decisions, or policy-makers' decisions. I demonstrate that methods that rely on suspect variation display the expected patterns of bias.

The methods I employ have the advantage that participants are not aware of being evaluated. In this way, the experiments mimic actual class size reduction policies, which rarely include evaluations or incentives for schools to make good use of the opportunities provided by smaller class sizes. If one were consistently to find that policy experiments that reduced class size *and* contained incentives produced greater improvements in test scores than natural experiments that just reduced class size, one might conclude that the incentive environment is important. That is, policies that just provide more resources may be significantly less efficacious than policies that link resources to performance.

APPENDIX TABLE

	Mean	Std. dev.	1st %ile	99th %ile
enrollment in grade 1	68.884	33.112	15	180
enrollment in grade 2	63.340	29.523	15	161
enrollment in grade 3	62.013	28.461	15	158
enrollment in grade 4	62.006	31.356	14	177
enrollment in grade 5	63.577	35.049	12	200
enrollment in grade 6	82.385	67.857	10	340
class size in grade 1	21.414	5.539	8	32
class size in grade 2	21.152	5.248	8	32
class size in grade 3	22.299	5.486	12	33
class size in grade 4	22.654	5.795	11	34
class size in grade 5	21.886	6.277	13	32
class size in grade 6	23.516	6.420	10	33
residual log enrollment in grade 1*	0.000	0.103	−0.707	1.586
residual log enrollment in grade 2*	0.000	0.106	−0.650	1.813
residual log enrollment in grade 3*	0.000	0.105	−0.696	0.947
residual log enrollment in grade 4*	0.000	0.107	−1.147	1.343
residual log enrollment in grade 5*	0.000	0.106	−0.727	0.519
residual log enrollment in grade 6*	0.000	0.096	−0.673	0.495
residual log kindergarten cohort*	0.000	0.080	−2.291	0.627
minimum class size grade 1	15.058	1.498	6	20
minimum class size grade 2	15.090	1.542	6	20
minimum class size grade 3	15.099	1.553	6	20
minimum class size grade 4	15.165	1.713	6	22

APPENDIX TABLE
(CONTINUED)

	Mean	Std. dev.	1st %ile	99th %ile
minimum class size grade 5	15.174	1.734	6	22
minimum class size grade 6	15.174	1.734	6	22
maximum class size grade 1	23.338	2.712	15	30
maximum class size grade 2	23.731	2.879	15	32
maximum class size grade 3	24.656	2.412	18	32
maximum class size grade 4	25.345	2.356	19	32
maximum class size grade 5	25.538	2.349	19	32
maximum class size grade 6	25.655	2.270	19	32
median household income	43,930	13,468	22,140	104,483
percent of the population who are African-American	7.368	10.723	0	41.184
percent of the population who are Hispanic	5.640	7.686	0	31.037
percent of adults who are at least high school graduates	83.579	6.985	62.566	95.304
percent of adults who are at least 4-year college graduates	28.066	11.874	7.836	60.768
percent of population who are urban	74.560	35.630	0	100

Enrollment in each grade is taken from the series variously titled *Strategic School Profiles, Town and School District Profiles,* and *Condition of Public Elementary and Secondary Education in Connecticut: Town and School District Profiles.* Class size for the school years 1991–1992 through 1997–1998 is taken from the same series. For previous years it is taken from unpublished data made available by the State of Connecticut Department of Education. In many cases class size has been checked against individual districts' annual reports. Class size has also been checked against the series *Elementary Classes, by Size, in Connecticut Public Schools.* The same series (*Strategic School Profiles* and so on) contains the demographic variables listed above (median household income through the percent of the population who are urban). The primary source for the demographic variables, however, is the *School District Data Book,* which is a school district level summary of the 1990 United States Census of Population and Housing. The size of each kindergarten cohort is taken from the series titled *Enumeration of Children* and from a similar series compiled by Claritas, Incorporated. For school level variables (enrollment, class size), there are 3504 observations in first grade, 3504 observations in second grade, 3464 observations in third grade, 3404 observations in fourth grade, 3071 observations in fifth grade, and 1150 observations in sixth grade. For district level variables (maximum and minimum class size rules, demographics), there are 1752 observations.

HARVARD UNIVERSITY AND
NATIONAL BUREAU OF ECONOMIC RESEARCH

REFERENCES

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

Angrist, Joshua, Guido Imbens, and Donald Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association,* XCI (1996), 444–455.

Betts, Julian, "Is There a Link between School Inputs and Earnings? Fresh Scrutiny of an Old Literature," in Gary Burtless, ed., *Does Money Matter? The Link between Schools, Student Achievement, and Adult Success* (Washington, DC: The Brookings Institution, 1995).

Card, David, and Alan Krueger, "School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina," *Journal of Economic Perspectives,* X (1996), 31–40.

Claritas, Incorporated, *Claritas Update,* Electronic data on population by age for towns in Connecticut (Arlington, VA: Claritas, Incorporated).

Connecticut Public Expenditure Council. *Elementary Classes, by Size, in Connecticut Public Schools* (Hartford, CT: Connecticut Public Expenditure Council, 1975 through 1988).

Ferguson, Ronald, "Can Schools Narrow the Black-White Test Score Gap?" in Christopher Jencks and Meredith Phillips, eds. *The Black-White Test Score Gap,* (Washington, DC: Brookings Institution Press, 1998).

Hanushek, Eric, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature,* XXIV (1986), 1141–1177.

——, "Measuring Investment in Education," *Journal of Economic Perspectives,* X (1996), 9–30.

Hanushek, Eric, John Kain, and Steven Rivkin, "Teachers, Schools, and Academic Achievement," National Bureau of Economic Research Working Paper No. 6691, 1998.

Harcourt-Brace Educational Measurement, *Connecticut Mastery Tests Interpretative Guide* (Connecticut State Board of Education: various years 1986 to 1998).

Hoxby, Caroline, "The Effect of Class Size and Composition on Student Achievement: New Evidence from Natural Population Variation," National Bureau of Economic Research Working Paper No. 6869, 1998.

Krueger, Alan, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics,* CXIV (1999), 497–532.

National Center for Education Statistics, *Digest of Education Statistics, 1998* (Washington, DC: Government Printing Office, 1999).

Salmon, Richard, Christina Dawson, Stephen Lawton, and Thomas Johns, *Public School Finance Programs of the United States and Canada,* 1993–94 edition (Denver, CO: American Education Finance Association, 1995).

State of Connecticut Board of Education, *Condition of Public Elementary and Secondary Education in Connecticut: Town and School District Profiles* (Hartford, CT: State of Connecticut Board of Education, 1977–1978 through 1981–1982).

State of Connecticut State Department of Education, *Enumeration of Children* (Hartford, CT: Connecticut State Printing Office, 1975 through 1983).

State of Connecticut Board of Education, *Strategic School Profiles* (Hartford, CT: State of Connecticut Board of Education, 1991–1992 through 1998–1999).

State of Connecticut Board of Education, *Town and School District Profiles* (Hartford, CT: State of Connecticut Board of Education, 1982–1983 through 1990–1991).

State of Connecticut Department of Education, Unpublished series on variables (enrollment, class size) that appear in recent (1993–1994 through 1998–1999) editions of *Strategic School Profiles* (Hartford, CT: State of Connecticut Department of Education, 1995).

United States Department of Education, National Center for Education Statistics, *School District Data Book: 1990 Census School District Special Tabulation,* Computer file (Washington, DC: National Center for Education Statistics, 1994).