NBER WORKING PAPER SERIES

# THE EFFECTS OF CLASS SIZE AND COMPOSITION ON STUDENT ACHIEVEMENT: NEW EVIDENCE FROM NATURAL POPULATION VARIATION

Caroline M. Hoxby

Working Paper 6869
http://www.nber.org/papers/w6869

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 1998

The author gratefully acknowledges the assistance of the Connecticut Department of Education, whose staff are not responsible for any statements or errors in the paper. The author acknowledges helpful ideas and suggestions from Henry Farber, Richard Freeman, 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 APPAM meetings. This research was ably assisted by Bridget Terry and Ilyana Kuziemko. The views expressed here are those of the author and do not reflect those of the National Bureau of Economic Research.

The Effects of Class Size and Composition on
Student Achievement: New Evidence from
Natural Population Variation
Caroline M. Hoxby
NBER Working Paper No. 6869
December 1998

## ABSTRACT

I use natural population variation to identify the effects of class size and composition on student achievement. I isolate the credibly random component of population variation in each grade and school district and use this component to generate instrumental variables for class size and composition. I also exploit the discontinuous changes in class size that occur when natural population variation triggers a change in the number of classes in a grade in a school. Discontinuity-based results are both consistent and precise only when applied to *within-district* changes in class size and population. I find that reductions in class size from a base of 15 to 30 students have no effect on student achievement. The estimates are precise enough to identify improvements in math, reading, or writing achievement of just 3/100ths of a standard deviation. I find that the presence of black students in a class, in an of itself, has no effect on achievement. I demonstrate that estimates of the effect of racial composition that rely on between-district comparisons suffer from substantial bias. Finally, I show that more female classes perform significantly better in writing in the 4th through 8th grades and in math in the 4th grade. Comparison of the effects to average male-female differences in test scores suggest that gender composition alters classroom conduct.

Caroline M. Hoxby
Department of Economics
Harvard University
Cambridge, MA 02138
and NBER
choxby@harvard.edu

## I. Introduction

In this study, I use a never-previously-used source of variation in school inputs, natural population variation, to identify the effects of class size and class composition on student achievement. I isolate the credibly random component of population variation in each school district and use this component to generate instrumental variables for class size and class composition. By using this source of variation, this study comes closer to a random experiment than previous methods. An added benefit is that natural population variation generates fluctuations in class size and composition that are in the range relevant to current policy debates. Moreover, unlike policy experiments with class size such as Tennessee's Project Star, 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 California's 1996 action, rarely contain built-in evaluations 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 because there is little or no debate over whether smaller classes provide better *opportunities* to improve achievement.[1] Debate focuses, instead, on whether smaller classes actually *do* improve achievement, given that schools need not take advantage of improved opportunities. In short, the evidence from natural population variation is directly relevant to policy questions about the efficacy of class size reductions.

When policy makers--local, state, or federal--want to improve student achievement, class size reduction is one of the tools they are most likely to use. In 1996, the California legislature dedicated 1 billion dollars to class size reduction. The administration's 1999 federal budget proposal contained 12 billion dollars (over 7 years) for the same purpose. Several states offer incentives for class size reductions in their school finance laws, and teachers' unions frequently ask for reductions in collective bargaining. Because the policy enjoys almost perennial popularity, class size has fallen monotonically in the United States during the twentieth century. As a result, elementary schools averaged 19 students per teacher in the 1997-98 school year.[2]

Another set of contested policies concern the racial and gender composition of classes. In recent years, the consensus that the ideal classroom is co-educational and racially and ethnically integrated has broken down. The breakdown reflects a scholarly debate over whether peer interactions in integrated, co-educational classrooms impede or enhance learning by females and members of minority groups.[3]

One of the key reasons that both class size and class composition policies are controversial is that the empirical evidence is contradictory.[4] The origin of the empirical controversy is essentially simple. Different studies use different sources of variation in class size and class composition, and nearly all of these sources of variation are potentially tainted by correlation with other determinants of student achievement. Very little of the variation in school inputs is effectively random with respect to student achievement and determinants of student achievement. The vast majority of variation in school inputs is the result of choices by parents, schooling providers, or courts and legislatures; and this variation 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 random unless there is some explicit reason why we should think it is.

This criticism does not apply to explicit experiments that have random variation built into their designs, and evidence from experiments like Project Star has manifest advantages.[5] These advantages are, however, offset by a few disadvantages. Explicit experiments are rare (tempting interpreters to extrapolate the results unduly), most experiments take place in developing countries (so that the range of inputs is not relevant for the United States), data are usually closely held (making it difficult to verify whether the experimental design was maintained), and--by far most important--the actors in the experiment are aware of it. The actors' awareness has several effects. First, being evaluated often makes people behave more productively for some time, even if the policy under evaluation has no effect (the "Hawthorne effect"). Second, people often try to undo the randomness of the design. Some principals may attempt to put children who they feel can benefit from more attention into the small classes. Other principals may assign their best teachers to the small classes or engage in extra monitoring of the small classes. Most importantly, the outcome of an explicit experiment generally determines whether the policy is continued and enacted universally. Thus, the actors have incentives to make the policy successful that they would not ordinarily have.

In this study, I use variation in class size and composition that comes from natural randomness in the population. The intuition is straightforward. Consider a school district that has a population that is in steady state. There is still natural variation in the timing and gender of births such that the entering kindergarten cohort varies somewhat in size, the ratio of females to males, and the ratio of minorities to non-minorities. This variation is not fully smoothed because there is discreteness in school entry rules (children born between date X and date Y must enroll in first grade in a particular school year) and because the number of classrooms in each school is an integer. In districts with big populations, the law of large numbers smooths natural population variation greatly. Also, a big district can keep class size relatively constant in the face of natural population variation by managing its teachers and classrooms flexibly. Thus, random population variation generates only a small amount of empirically useful variation in class size and composition in large districts. However, in small districts--at the extreme, one classroom per grade--natural population variation translates almost directly into differences in class size and class composition between cohorts. In a small district that is in steady state, one cohort might persistently experience classes that are small and two-thirds female in grades one through six, while the subsequent cohort might persistently experience classes that are large and half female. The different cohorts are essentially randomly assigned different treatments.

I attempt to isolate the random component of population variation using a long panel of data on school-age population, enrollment, and births in Connecticut school districts. The panel allows me to eliminate nearly all smooth changes in population: I use just residual population changes that remain after fitting a quartic function of time *separately for each district*. Using the data on births, I confirm the results based on enrollment and eliminate any possibility that the population variation being used is endogenous to revealed class size. I also identify the class size effect using the fact that class size jumps abruptly in small schools when a class is added to or subtracted from a grade. Such discontinuities in the relationship between class size and enrollment make natural population variation provide another, independent source of identification--that is, overidentification–that can be used to test the main results.

The one disadvantage of random, natural population variation is that it is necessarily transitory from the individual teacher's point of view.[6] I discuss this point in interpreting the results on class size.

## **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 parental choice is likely to generate upward biased estimates of the efficacy of inputs. The same may be said for systematic variation within a district a over time. For instance, class size reductions will be appear to be more efficacious than they really are if the types of parents who contribute more to their children's learning choose school districts that offer smaller class sizes.

Even if we identify parents who have similar attributes, there is ample potential for bias. Parents choose school inputs endogenously, based on their children'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 similar parents, 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. If, on the other hand, providers attend more to children with learning problems, estimates will be downward biased.

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.[7] The negative bias resulting from the use of compensatory policies is, however, often offset by positive omitted variables bias caused by policy-makers' simultaneous pursuit of compensatory labor market policies. For example, policies that decreased racial discrimination in school inputs were implemented simultaneously with policies that decreased racial discrimination in employment.

In short, it is not surprising that the results of empirical studies differ (that is, suffer from different biases) depending on the source of the 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 and isolating a source of random variation is preferable to simply eliminating all obvious sources of bias. This point is illustrated by the debate over Card and Krueger's [1992] paper. They exploit the fact that, within each of the nine U.S. census regions, residents have experienced a variety of school input levels because they were schooled in a variety of states at a variety of times. Heckman, Layne-Ferrar, and Todd [1995] show that the variation used is partly due to selective migration within the United States: able people from states with low returns to skill move to states with high returns to skills. Such subtle paths for omitted variables bias are easily overlooked.

## III. Empirical Strategy

In the exposition that follows, I consider class size as the school input, but the empirical strategy works equally well for class composition. Consider the achievement of students in grade *i* of school *j* in district *k* in year *t*. It is determined by class size as well as unobserved attributes like student ability and parental contributions to learning. We typically estimate an equation such as:

$$A_{ijkt} = \beta_1 C_{ijkt} + X_{ijkt}\beta_2 + \epsilon_{ijkt} \qquad (1)$$

where $A$ is achievement; $C$ is class size; $X$ is a vector of observed student, parent, and community characteristics; and $\epsilon$ is all other determinants of achievement, including the unobserved attributes of the students, parents, and community. By definition, class size is equal to regular enrollment divided by the number of classes (properly defined). For all the reasons discussed above, the number of classes, which is a policy variable, is a function of student, parent, and community characteristics, observed and unobserved. Enrollment is partly a function of these characteristics, but it also depends on random variation in the population of children who are in the age range appropriate for a given grade in a given year. Formally,

$$C_{ijkt} = \frac{E_{ijkt}}{n_{ijkt}(X_{ijkt},\epsilon_{ijkt},E_{ijkt})} = \frac{\check{E}_{ijkt}(X_{ijkt},\epsilon_{ijkt}) \cdot u_{ijkt}}{\tilde{n}_{ijkt}(X_{ijkt},\epsilon_{ijkt},u_{ijkt})} \tag{2}$$

where $E$ is enrollment and $n$ is the number of classes. Ě is what enrollment would be if the timing of births were deterministic, rather than random, functions of a population's characteristics. For instance, if parents could and did time all births precisely, then enrollment would be deterministic and actual enrollment would equal Ě. $u$ is the variation in enrollment that results from the fact that biology causes random variation in the timing of births. $E$ and Ě are functions of student, parent, and community characteristics ($X$ and $\epsilon$). $n$ is not only a function of $X$ and $\epsilon$, but also of enrollment. That is, the costs and benefits of adding another class depend not only on how much local parents care about schooling but also on the actual enrollment in any given year, which is partly random. Thus, the denominator on the far right-hand-side of equation (2) shows that $n$ is ultimately a function of $X$, $\epsilon$, and $u$. $u$ affects Ě proportionally, so that equation (2) works for populations of different sizes.

### 1. The First Identification Method

$u$ is not correlated with $X$ and $\epsilon$, which are determinants of achievement, but $u$ is a determinant of both $E$ and $n$, so it is a good instrument for class size. I attempt to isolate $u$ using the fact that the aggregate characteristics that affect achievement change much more continuously than enrollment does for a particular grade in a particular school in a particular year. For instance, while school entry dates make small age differences between children potentially cause large differences in the cohort that belongs to each grade-year, small age differences between adults in a community have only small effects on achievement. Thus, while Ě is not static, it does not change abruptly.

We can characterize *any* Ě that changes smoothly over time by a grade-school-district-specific intercept and a grade-school-district-specific polynomial in time. (This statement holds equally for the log of Ě.) That is, we can write:

$$\ln(\breve{E}_{ijkt}) = \alpha 0_{ijk} + \alpha 1_{ijk} t + \alpha 2_{ijk} t^{2} + \alpha 3_{ijk} t^{3} + \alpha 4_{ijk} t^{4} + \ldots \qquad \textit{or}$$

$$\ln(E_{ijkt}) = \alpha 0_{ijk} + \alpha 1_{ijk} t + \alpha 2_{ijk} t^{2} + \alpha 3_{ijk} t^{3} + \alpha 4_{ijk} t^{4} + \ldots + \ln(u_{ijkt}) \tag{3}$$

I estimate such an equation separately for each grade-school-district.[8] 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. The estimated residual gives us a consistent estimate of *u*, which is a good instrument for class size.

The method just described exploits the fact that the aggregate characteristics that affect achievement, *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 classes are unusually large for his grade-school-district. Even if the cause of the large classes is random population variation, the parent might decide to send his child to a private school or attempt to have his child held back a grade or moved up a grade. Such reactions, although probably rare, would have the potential to make *X* and ϵ endogenous to *u*. Fortunately, a simple modification of the method eliminates this problem.

Rather than use enrollment in a grade-school-district as the dependent variable in equation (3), one can use the population of children predicted to be in a grade-school-district based on births that occurred in the district such that, if the babies grew up and enrolled according to the birth date rule, they would be in the grade.

I call this variable "births-in-grade" and represent it by *B*. To summarize, the estimated residual from

$$\text{(4)} \qquad \ln(B_{ijkt}) = \tilde{\alpha}0_{ijk} + \tilde{\alpha}1_{ijk}t + \tilde{\alpha}2_{ijk}t^2 + \tilde{\alpha}3_{ijk}t^3 + \tilde{\alpha}4_{ijk}t^4 + \ldots + \ln(u_{ijkt})$$

gives us a credibly consistent estimator for *u* that has no potential to be correlated with *X* or $\epsilon$ through their possible endogeneity to realized class size.

If schools never added or subtracted classes, then we would expect to find a coefficient equal to one if *ln(C)* were regressed on an estimate of *ln(u)*. Random population variation can, however, cause schools to change *n*, so the relationship between *ln(C)* and *ln(u)* will not be linear with a coefficient equal to one. In the first identification method , I treat changes in *n* as nuisances that happen to cause non-linearities in the relationship between *ln(C)* and *ln(u)*, and I rely on the exogeneity of *ln(u)* for identification. (The second identification method relies solely on the discontinuities in the relationship between class size and enrollment that are generated by changes in *n*.)

Observe that we can write class size as:

$$\text{(5)} \qquad C_{ijkt} = \frac{E_{ijkt}}{n_{ijkt}} = \frac{E_{ijkt}}{n_{ijk,t-1} + I^{add}_{ijkt} - I^{subtract}_{ijkt}}$$

where $n_{ijk,t-1}$ is the number of classes in the previous year, $I^{add}$ is an indicator variable for adding a class, and $I^{subtract}$ is an indicator variable for subtracting a class. I do not assume that adding and subtracting occur symmetrically because adjustment costs are not symmetric (particularly because the vast majority of teachers have tenure).

Because I have panel data and actually observe each addition or subtraction of a class, I can estimate probit equations for adding a class and subtracting a class. For instance, I can estimate:

$$\text{(6)} \qquad Prob(I^{add}_{ijkt}{=}1) = Prob(n_{ijkt}{=}n_{ijk,t-1}{+}1) = \Phi\left(\delta_{0k} + \delta_{1k}\frac{E_{ijkt}}{n_{ijk,t-1}} + \delta_{2k}\left(\frac{E_{ijkt}}{n_{ijk,t-1}}\right)^2 + \ldots + \delta_{8k}\left(\frac{E_{ijkt}}{n_{ijk,t-1}}\right)^8\right)$$

for each district, where the high-order polynomial specification is used to pick up nonlinearities.[9] In the second identification method, I estimate probit equations and plug in enrollment. But, in the first identification method, where only the variation in enrollment that is plausibly natural population variation is used, I would not plug in actual enrollment but estimates of enrollment based on *u*. The results of this exercise would then be used as instruments for class size, in addition to *u*. Calculating correct standard errors under this elaborate procedure would be extremely difficult, so the practical thing to do in the first identification method is not to estimate equations for $I^{add}$ and $I^{subtract}$ at all, but to skip this intermediate step and instrument for class size with not just *ln(u)* but functions of *ln(u)* as well.[10] I use powers of *ln(u)* up to the 8th power (these instruments have ample explanatory power), but other functions of *ln(u)* could be used too. Most of the explanatory power in the first stage regression comes from the linear relationship between class size and *ln(u)* that exists when the number of classes does not change.

In short, my first identification method estimates equation (1), which relates achievement to class size, by instrumenting for class size with an estimate of *ln(u)*and powers of the estimate of *ln(u)*.[11]

2. The Second Identification Method

The second identification method 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 take advantage of these discontinuities is the cross-section method of exploiting maximum class size thresholds. Angrist and Lavy [1997] illustrate this method using Israeli schools, which are supposed to have a maximum class size of 40. In America, districts have maximum class sizes that are much lower. For instance, if a school had a maximum class size threshold of 24, it would put students into one class until enrollment is 24, put students into two classes until enrollment is 48, and so on. Its rule could be written:

$$(7) \qquad C_{ijkt} = \frac{E_{ijkt}}{int\left(\dfrac{E_{ijkt}-1}{C^{\max}}\right)+1} \quad where\ C^{\max}=24\ and\ int(z) \equiv greatest\ integer\ \leq\ z$$

and its relationship between class size and enrollment would be illustrated by Figure I. Notice that class size varies abruptly and predictably when enrollment is at a multiple of 24. These discontinuities provide identification because the difference in the underlying population that produces enrollment of 24 versus 25 is very small (and should have a correspondingly small effect on achievement), but the difference in class size for enrollment of 24 versus 25 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 24 and enrollment of 25 based solely on the rule given by equation (7) is a good instrument for the actual difference in class sizes between schools with enrollment of 24 and 25. The same is true for 48 and 49, 72 and 73, 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 *ln(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 $X$ and $\epsilon$. 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 (7). All variation in predicted class size that is not generated by a rule-triggered change in the number of classes 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 useful variation is the variation inside in the narrow ranges enclosed by the dotted line in Figure I--that is, the difference in achievement for enrollment of 24 and 25, for enrollment of 48 and 49, *et cetera*. In cross section data, all of the other variation in enrollment is suspect because it is between-district variation that not only reflects differences in the underlying populations ($X$ and $\epsilon$), but is plausibly endogenous to realizations of class size. (Some schools *routinely* have larger class sizes than others because of the way the rule functions, and parents 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 in very narrow ranges. Angrist and Lavy do only some of the proper discarding because their cross-section data contain too few occurrences of enrollment in the right ranges. As a consequence, they sacrifice consistency for power. Below, I present cross-section method that demonstrate what happens when proper discarding is not done. Since my data are actually panel data, I am able to employ a within-district method (described below) that is both 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, some American states have maximum class size rules that do not bind in districts where households have above-average incomes. Similarly, in Angrist and Lavy's data, the maximum class size rule does not bind in districts that serve well-off households. There is nothing wrong with estimating the effect of class size only for less-well-off students, but one must be careful to interpret the results correctly. Also, if better-off districts actually have maximum class size rules of their own that they follow, then using a state-wide or country-wide rule that does not bind everywhere is throwing away good variation. Since there is typically not much good variation to exploit in discontinuity-based identification strategies, throwing it away is undesirable.

The within-district method, which produces results that are both more representative and more powerful, starts with consistent estimation of each district's rules for adding and subtracting a class. I have data on multiple grades in multiple schools in multiple years in each district. Thus, I estimate two probit equations *for each district*, where the dependent variables are $I^{add}$ and $I^{subtract}$ and the specification is that given by equation (6).[12] Since I actually observe each time a class is added to a grade in a school, I can let the probability of adding or subtracting a class be state-dependent. That is, equation (6) allows for the fact that a school with 51 students in a grade is more likely to divide them into 3 classes if it used 3 teachers in that grade in the previous year than if than if used only 2 teachers.[13]

Armed with estimates of each district's add-a-class and subtract-a-class rules, I declare the add-a-class threshold to be the class size at which the probability of adding a class exceeds 50 percent and the subtract-a-class threshold to be the class size at which the probability of subtracting a class exceeds 50 percent. If, by chance, the wrong thresholds are selected for a district, the only result is that useful variation is thrown away.[14] Using the wrong thresholds is like using a rule that does not bind for a district.

In the data, the estimated add-a-class thresholds range from 22 to 40, with most thresholds falling between 24 and 28. The estimated subtract-a-class thresholds range from 8 to 18, but the vast majority of these thresholds are in the small range between 14 and 17.

I use the thresholds and actual enrollment data to calculate the difference between what class size would be if the number of classes did not change and what class size would be if the predicted changes in the number of classes occurred:

$$\frac{E_{ijkt}}{n_{ijk,t-1}+\hat{I}^{add}_{ijkt}-\hat{I}^{subtract}_{ijkt}} - \frac{E_{ijkt}}{n_{ijk,t-1}} \tag{8}$$

$$where\ \ \hat{I}^{add}_{ijkt}=1\ \ if\ \ \frac{E_{ijkt}}{n_{ijk,t-1}}>C^{\max};\ 0\ \ otherwise \qquad \hat{I}^{subtract}_{ijkt}=1\ \ if\ \ \frac{E_{ijkt}}{n_{ijk,t-1}}<C^{\min};\ 0\ \ otherwise\ \ .$$

(8) is the part of the change in class size that is due solely to the triggering of an add-a-class or subtract-a-class rule.[15] I discard observations for which (8) is zero, and use the remaining observations to instrument for the change in class size between this year and the previous year. That is, I estimate equation (1) in first-differences: the change in achievement is regressed on the change in class size, instrumented by the change in predicted class size that arises solely because a rule is triggered.[16]

## IV. Data

The empirical strategy creates a number of data requirements. First, since school cohorts are defined by birth date, the data must contain information on birth by calendar year or population-by-age on the legal cut-off (which is January 1 in the state of Connecticut).[17] Second, the data must contain districts that are not large because, in large districts, natural population variation averages out to a great extent within each cohort. Third, 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 much 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. 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 resulting emphasis on elementary class size fits the empirical and pedagogical debates, which are focused on class size in early grades.

Connecticut school data are particularly appropriate for the empirical strategy. The state has 163 elementary school districts, many of which are small. Half have typical cohort sizes smaller than 150 students. Districts are towns in Connecticut, so that vital statistics provide births by district of residence, by race of the baby.[18] Also, Connecticut collects an annual Enumeration of Children: population-by-age data as of January 1 for all school-aged children, by town. These data are unusual: most states rely on the decennial census. Finally, every year since 1979, Connecticut has administered state-wide achievement tests in mathematics, language arts, reading, and writing. Between 1979 and 1985, the tests were administered only in the ninth grade. Ninth grade scores are not very suitable for examining the effects of elementary class size and composition, so I make little use of these scores beyond checking that they confirm other results. In 1986, tests began to be administered in the fourth, sixth, and eighth grades. I mainly use these eleven years of test data to measure achievement.[19] In all these years, class size in Connecticut averaged 20 students with a standard deviation of 3.6 students, but classes are observed that are as large as 40 students and as small as 8.[20] 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, children who enter 6th grade in the fall of 1991 are in the graduating class of 1998. All the data are available by district by grade. Enrollment (by race and gender) and class size data are available by school by grade. I have 11 years of achievement data for each grade, so each achievement equation contains 1793 observations (163 districts times 11 years of panel data on a single grade). I have 23 years of enrollment data for each grade, however, so I estimate the enrollment residuals based on all 23 years of data. The larger number of years allows me to get more precise estimates of $u$.[21]

It is worth considering what the correct unit of analysis would be if all the data were available by school. If school attendance area boundaries were rigid (not in the control of the district or only changed with great difficulty) and if transfers between schools within districts were never allowed, then the district would be unable to control natural population variation at the school level and the school would be the best unit of analysis. That is, school-level analysis would be most powerful, though aggregating the data to the district level and proceeding with district-level analysis would produce unbiased results. District-level analysis is best if the district can exercise some control over natural population variation at the school level. For instance, suppose a district finds that one of its schools is facing an unusually large kindergarten cohort. The district might reduce the cohort's size by revising its school attendance areas or by offering a popular program, such as full-day kindergarten, at another school to induce voluntary transfers. District-level analysis would be unbiased in this case, but school-level analysis might be biased.

In practice, the achievement equations must be run at the district level since confidentiality considerations prevent the release of school-level test scores. One can, however, estimate the enrollment residuals (using equations 4 and 5) and the first-stage equations at the school level. For the first identification method, I found that school-level estimation generated an increase in power over district-level estimation that was too small to justify the possibility of bias that school-level estimation introduced. For the second identification method, the add-a-class and subtract-a-class probit equations *should* be estimated at the school level because they are attempting to identify a district's rules. That is, they are not attempting to exploit variation beyond a district's control, but to understand how a district controls itself.

The tests are administered at the beginning of each school year (September). Thus, the 4th grade tests may be affected by class sizes in the 1st through 3rd grades, but they are unlikely to be affected by 4th grade class size. Similarly, class sizes in 1st through 5th grades are relevant for the 6th grade tests, and class sizes in 1st through 6th grades are relevant for the 8th grade tests. I have attempted to test these statements of timing relevance in the data, and they are generally confirmed--although the tests are weak because most cohorts experience similar class sizes in the 1st through 6th grades. Unusually large cohorts, for instance, tend to consistently experience large class sizes. In other words, although I match tests to the relevant class sizes, the timing of the matches is not actually crucial to the results.

A class is defined as a group of students who spend the majority of the school day under the supervision of one teacher. The measure of class size excludes pull-out instruction by special education teachers or aides. No teachers' aides for regular instruction were observed in the districts that provide the useful variation--that is, outside of the largest districts in the state. Similarly, mixed-grade classes were so rare in the districts that provide useful variation that I could not estimate an equation that predicted when they would be introduced.[22] Instead, I include an indicator for mixed-grade classes.

All the data used are from publicly available sources.

## V. Some Illustrative Graphs

Graphs for individual school districts can provide intuition about the empirical strategy and the results. I consider three school districts in Connecticut, chosen for their illustrative value (rather than their representativeness): an extremely small district, a very small district, and one of the 10 largest districts in the state. Each of Figures IIa through IIc shows a district's enrollment and class size in the 4th grade, by cohort. I selected the fourth grade because test scores are first available in that grade, but it would not have mattered much if I had selected another grade. Figure IIa shows that, in the extremely small district, enrollment and class size are identically equal for every cohort. The district has only one classroom per grade. In the small school district (Figure IIb), class size varies very closely with cohort size except for the graduating class of 2001. In this cohort, enrollment was 57 students and, instead of dividing the grade into two classes as was usual, the school put the cohort into three classes of 19 students. This figure thus illustrates the abrupt difference in class size that can occur when a change in the number of classes is triggered in a small school. In both the small districts, class size varies over a large range: the smallest class observed contained 11 students, the largest 31 students. This range covers the policy range very fully--in fact, more fully than commonly used longitudinal surveys like High School and Beyond or experiments like Project Star. In the large district (Figure IIc), there is relatively little variation in class size because its large enrollment smooths variation and its large number of classrooms and teachers makes flexible management easy.

To measure classroom racial and gender composition, I use exposure or the average weighted percentage of classes that are black or female in a grade-district. In Figures IIIa through IIIc, I show the average weighted percentage black in 4th grade for each cohort for the same three school districts. Figures IVa through IVc show similar series for the average weighted percentage female. The purpose of Figures IIIa through IVc is to demonstrate how much more variation in exposure exists in small districts than in large districts. In the extremely small district, the black percentage varies between 0 and 17 percent--without showing any trend. In the same district, the female percentage varies between 33 percent and 58 percent. In contrast, in the large district, the black percentage varies only between 21 percent and 26 percent and the female percentage varies only between 46 percent and 51 percent.

Figures Va through Vb show math scores for the same three districts, with class size graphed also. If reducing class size improved math scores, then we would expect to see the two lines generally move in opposite directions, like a mirror images of one another. But, it is difficult to discern any pattern linking math scores and class size. The same can be said for Figures VIa through VIc, which show reading scores and class size, and for Figures VIIa through VIIc, which show writing scores. However, looking at these three districts 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.

## V. Class Size Results

In this section, I examine the effects of class size on achievement. For comparison, I show results for not only the appropriate estimation method, but also less appropriate (though common) methods.

Table II introduces the class size results by showing what happens as we move from less to more appropriate estimation methods. The dependent variables are formed by dividing each test score by the overall standard deviation of scores on that test in Connecticut. Thus, the estimates in the table show how test scores, measured in standard deviations, change when class size changes by one student. Dividing by the standard deviation puts all the test scores into a convenient and easily interpreted metric, especially since the raw scores are not intuitive, especially for reading and writing. It may useful, however, to know that a standard deviation on the math test corresponds to mastering 2 or 3 more objectives out of a total of 40.

In addition, I estimate value-added specifications using the differences between 6th and 4th grade scores and between 8th and 6th grade scores for a given cohort in a given district.[23] These specifications allow me to determine whether class size affects not only the level but the rate of growth of student achievement.

Each cell in Table II shows the estimated coefficient on class size *from a separate regression*. The specification of each regression is described by the column and row headings. For instance, the number in the upper-left-hand cell is the effect of average class size in grades 1 through 3 on 4th grade math scores using a specification that pools observations across districts and cohorts (with cohort fixed effects). This specification is a naive one likely to produce estimates biased by correlation between class size and unobserved parent and community attributes. In fact, the estimates are all negative and highly statistically significant. If we were to give them credence, we would interpret them as indicating that a one student reduction in class size improves test scores by about 1/10th of a standard deviation.

In column II, I add demographic variables such as median household income, the percentage of the population in poverty, the percentage of adults who are college graduates, and the percentages of the population who are black and Hispanic. Also included is an indicator for the existence of a two-grade class. These simple controls for observed parents and community characteristics greatly attenuate the estimated effect of class size on test scores, but the results are still mostly negative in sign. Five of the fifteen estimates are statistically significant at the 5 percent level; all five are negative. In fact, the results shown in column II represent a familiar pattern of results for between-school comparisons that control for simple demographics.

Column III includes district fixed effects, which control for any district characteristics--observed or unobserved--that are constant over time. We cannot sign the bias we expect in such a regression, since much of the identifying variation in class size now comes from within-district decisions to reduce class size over time or reduce class size for certain students. In fact, the statistically significant estimates are mixed in sign.

The final column of Table II includes not only district fixed effects but also district-specific linear time trends. These results are somewhat credible since a lot of the non-random variation in class size is partialed out by linear time trends. However, the statistically significant estimates are still mixed in sign. Notice that the within-district specifications (columns III and IV) generate very small standard errors: a change as small as 1/100th of a standard deviation would generally be statistically significant at the 5 percent level. These small standard errors are indicative of the fact that, after the addition of district fixed effects and district-specific linear time trends, a large amount (about 75 percent) of the variation in class size remains.

### 1. Results from the First Identification Method

Table III shows the explanatory power of the instruments in the implied first-stage equations. Each cell represents a separate regression, and each contains the partial F-statistic on the joint significance of the excluded instruments (*ln(u)* and powers of *ln(u)*). Each column heading describes the method of estimating *ln(u)* being used. For instance, column IV uses powers of the residual of enrollment from district-specific regressions that contain an intercept and a quartic in time.

The F-statistics indicate that the excluded instruments are powerful predictors of class size. The F-statistics in columns I through IV, which use enrollment residuals, are almost uniformly greater than 10. There is no clear difference in explanatory power between residuals from a cubic and residuals from a quartic, suggesting that higher order polynomials would not explain more of the smooth changes in enrollment than a quartic does. The quartic is, therefore, my preferred specification.

Column V contains similar estimates for residuals from births-in-grade. As we expectd, these residuals are powerful instruments, especially for later grades. By 6th grade, residential mobility provides ample room for spillage between births and enrollment. Nevertheless, births-in-grade residuals are sufficiently powerful to generate a specification test for results based on enrollment-in-grade residuals.

Table IV contains the main class size results for the first identification method. Again, each cell contains an estimate from a separate regression. Columns I through IV use different enrollment-in-grade residuals as the instrument for class size. Column V uses births-in-grade residuals.

Before considering the estimated coefficients, note the standard errors. The standard errors are larger with instrumental variables than they were with ordinary least squares in Table II. Nevertheless, they are small enough that changes in test scores as small as 2/100$^{\text{ths}}$ to 4/100$^{\text{ths}}$ of a standard deviation would generally be statistically significant at the 5 percent level in columns I through IV. In other words, if reducing class size by 1 made students master 0.1 more objectives (out of 40) on the math test, 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 contain no evidence of achievement gains linked to smaller class sizes. The estimates are mixed in sign, and none is statistically significant at the 5 percent level. (The few estimates that would be marginally significant at the 10 percent level have the "wrong" sign). In fact, given the standard errors, the effects are rather precisely estimated to be close to zero.

Column V presents estimates that use residuals from births-in-grade. The standard errors are considerably higher (although changes as small as 1/10$^{\text{th}}$ of a standard deviation would be statistically significant at the 5 percent level for most test scores). The point estimates do not, however, suggest a different pattern than those in columns I through IV. This suggests that few parents endogenously change their enrollment decisions by, say, sending their child to private school when they observe their child in an unusually large cohort.

Given that Table IV presents "well-estimated zeros," it is tempting 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. Table V contains those most likely to be of interest.[24]  Column I uses class size in the most recent grade as the measure of class size (whereas Table IV used average class size experienced in the relevant interval).  The results may differ little from those Table IV because a cohort tends to experience the same class sizes in 1st through 6th grade as a result of natural population variation.  In fact, we find little difference in the results. Columns II through IV use class size in the early grades, in order to test the hypothesis that class size reductions are more efficacious in early grades.  This hypothesis is prompted partly by Project Star results, which suggest that achievement gains are a one-time (level, not rate-of-gain) response to one early year of small class size.   The results in Columns II through IV are not, however, distinguishable from the main results in Table IV.

Column V shows the results of estimating a spline in class size at 23.  The spline is one of several tests of decreasing marginal returns to class size reductions that I estimated.  The threshold at 23 is motivated by Ferguson [1997], who argues that reductions in class size have no effect below 23 students, but that reductions above this threshold matter.[25]  Each cell shows the coefficient on class size when it is greater than 23.  The results shown are rather precisely estimated zeros: reductions in class size do not appear to be more efficacious above 23.

The next column also looks for decreasing marginal returns.  The independent variable is an indicator for the cohort's having experienced class size of 30 or greater in at least two out of the three most recent grades. The coefficients and standard errors are, of course, different because the independent variable is constructed differently, but the general pattern of results does not differ substantively from that of the main results.

Finally, column VII uses an indicator for the cohort's having experienced class size of 15 or fewer in at least two of the three most recent grades.  This specification is motivated by the idea that some types of instruction might be practical only in very small classes.  None of the estimated coefficients is, however, statistically significantly different from zero at the 10 percent level.

Of course, it is not possible to test the efficacy of reducing class size in ranges that are not observed in the data reasonably often. It would be a mistake to extrapolate these results to schools in which class size is typically higher than 30. Since most schools in countries that are not highly industrialized fall into this category, the results cannot confirm or contradict most developing country studies. It would also be a mistake to extrapolate these results to class sizes of less than 15. In any case, such tiny classes are too expensive for most American districts to consider because the cost of a one student reduction accelerates as class size gets smaller (cost is linear in the percentage reduction, not the student reduction). 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.

Reductions in class size may be more efficacious in schools where children come from disadvantaged families. The estimates presented in Table VI attempt to test this hypothesis. The class size variable is interacted with a set of indicator variables for (i) household income in the district (left-hand panel), (ii) the poverty rate in the district (middle panel), and (iii) the percentage of the population in the district who are black (right-hand panel). Both rural and urban districts in Connecticut have concentrations of low income households, but poverty and, even more, black households are concentrated in urban districts. A value for income, poverty, or percent black at or below the 25th percentile puts a district in the "low" category. A value at or above the 75th percentile puts a district in the "high" category, and the remaining districts are in the "medium" category. The standard errors are slightly higher than in the main results because the sample identifying each coefficient is smaller. Fortunately, for the disadvantaged groups, effects as small as 4/100ths of a standard deviation would always be statistically significant at the 5 percent level for the 4th and 6th grade tests and the value-added specifications.

Nevertheless, none of the estimates is statistically significantly different from zero. Even the pattern of point estimates provides no confirmation of the hypothesis that class size reductions are more efficacious in districts that contain disadvantaged students.

### 2. Results from the Second Identification Method

Table VII shows results identified by the discontinuities that arise when the number of classes is changed. As with the first identification method, I attempt to show how the estimates change as the method applied becomes more appropriate. In columns I through IV of Table VII, I treat the data as though it were cross-section data, estimate the predicted class size function for each district based on its maximum class size and equation (7), and use the predicted function as an instrument for class size. In column I, I use the entirety of the function. Column II uses observations within 8 students of a discontinuity, and column III uses observations within 4 students of a discontinuity. Column IV uses only the discontinuities and is, of course, the desired specification that should produce consistent estimates. The estimates in columns I though III are likely to be biased, with the degree of bias being highest in column I.

The notes to Table VII contain information on the number of effective observations in each regression and the power of each first-stage equation. As one narrows in on the discontinuities, the number of usable observations falls from 1793 in column I to about 115 in column IV (the exact number depends on the grade). There is ample explanatory power in the first-stage regressions for column I, but the explanatory power is just adequate in the first-stage regressions for column IV: the t-statistics on predicted class size range between 2.0 and 2.7. These numbers demonstrate the extraordinary demands that the cross-section method puts on data when it is applied appropriately. The within-district method allows us to exploit more true discontinuities. The number of usable observations in column V is about 650 (the exact number depends on the grade).[26]

If we were to naively interpret the cross-section results in column I, we would conclude that a one student reduction in class size raises math and reading achievement by about 1/10th of a standard deviation. These results are roughly in line with those in column I of Table II, which shows the results of naive OLS regression. As we narrow in on the discontinuities (still using the cross-section method), statistical significance falls. But, even in column III (which uses observations within 4 students of a discontinuity), five of the estimates are statistically significant at the 5 percent level. In column IV, however, where only the discontinuities are used, none of the results is close to being statistically significant and the point estimates do not exhibit a pattern even in their signs. Therefore, the statistically significant results in the other cross-section results rely *not* on the discontinuities in the predicted class size function, but on the suspect parts of the function. Columns I through IV should make us skeptical of cross-section results that do not rely solely on discontinuities.

Column V shows results from the within-district method, in which a district's change in scores is regressed on its change in class size, instrumented by part of the change in predicted class size that is due solely to a maximum or minimum threshold being triggered. See equation (8). Like the column IV estimates, the column V estimates are consistently estimated. Unlike the column IV estimates, however, the column V estimates have standard errors that would allow a change in test scores of 3/100ths to 5/100ths of a standard deviation to be statistically significant at the 5 percent level. None of the coefficient estimates in column V is, however, statistically significant. The estimates are close to zero and mixed in sign, much like the estimates that depend on the first identification method.

## 3. Interpretation

Results from both identification methods indicate that class size reductions in the policy range have little or no effect on achievement. The estimates are precise enough so that improvements that are educationally significant would be identifiable. The two identification methods are checks on one another, and the results are robust to numerous specification changes.

How might we interpret these results, especially in light of the Project Star results, which show effects of class size reductions that are small but statistically significant (and would be statistically significant if they appeared in this paper, given this paper's standard errors)? In both the natural and policy experiments, teachers had more *opportunity* to improve achievement with smaller classes The difference between the two types of experiments may be that the natural experiment varied opportunities but did not vary incentives, while the policy experiment combined greater opportunities with evaluation and incentives for teachers and administrators to use the opportunities. If this is the correct interpretation of the difference in the results, then the implication is that policies like California's should contain built-in evaluation and incentives. There are also less positive interpretations of the difference between the natural and policy experiment results. It may be that the policy experiment works differently because of Hawthorne effects (especially since Project Star witnessed a one-time improvement in achievement, not a permanently higher growth rate of achievement). Or, it may be that school staff undid some of the random design of the policy experiment. Since Connecticut school staff were certainly unaware of the natural experiment, we do not have to be concerned about their responding to the evaluation.

We might attribute some of the difference in results to the necessarily transitory nature of natural population variation. The variation is transitory from the teacher's point of view, not from the student's or parents' point of view (since elementary class size is persistent for a cohort in a district). Teachers may be ill-prepared 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 he 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 any teacher's basic repertoire: answering questions, correcting assignments, tutoring after school, talking to parents, and so on. Also, the Project Star results were achieved after only one year of smaller class size, and the teachers involved did not receive special instruction.

## VI. The Effects of Racial Composition

Racial composition fits easily into the framework set up for class size. In particular, racial composition is likely to be correlated with unobserved student, parent, and community characteristics, generating the potential for bias. Thus, we can substitute percent-black-in-enrollment and percent-black-in-births in equations (3) and (4) and proceed to estimate residuals as outlined above.

Although the same empirical strategy works for racial composition, the question we want to answer is different than the class size question. Class size is largely a policy choice, and the policy range is clearly defined and similar across the United States. When racial composition is a policy choice, as under a desegregation plan, the policy range is different for every district. Desegregation plans generally attempt to equalize exposure over all classes in a district, but equal exposure is 15 percent exposure in some districts and 80 percent in others. Moreover, desegregation orders are responsible for only a small amount of school racial composition. Most racial composition is the result of families' choices about where to live (which may be affected by how other families or local schools treat their children). In short, racial composition is not generally a policy choice.[27]

The problem a family or school usually faces is not how to deal with a desegregation order, but how to deal with the fact that a black student has enrolled. In particular, families and schools might engage in faulty statistical discrimination: attributing traits they associate with blacks to the student and reacting by leaving the local schools, discouraging the black student, or causing contention in the classroom. I say "faulty" statistical discrimination since the fact that the student has enrolled indicates that he is likely to have unobserved traits that differ from black students who have not enrolled. Thus, the question is, what should families and schools expect of a black student when they observe him enroll? Should they expect him to have lower achievement than a non-black student who enrolls? Should they expect his presence to lower the achievement of non-black students?

The estimates I present help to answer these questions because natural population variation makes exposure vary randomly within a given steady state. In other words, natural population variations helps us isolate the effects of black students on average achievement that are *intrinsic* to their presence from effects that caused by other factors correlated with their presence. Examples of intrinsic effects would be blacks always having low innate ability or always causing other students' achievement to suffer.

Table VIII shows what happens as we move from less appropriate estimation methods to more appropriate ones. Each cell in the table shows the estimated coefficient on percent-black-in-class from a separate regression. The specification in column I simply pools observations across districts and cohorts (with cohort fixed effects). This naive specification is likely to produce estimates biased by correlation between percent black and unobserved parent and community attributes. In fact, all the estimates in column I are negative and highly statistically significant (seven of the t-statistics exceed 40). If we were to give the results a causal interpretation, we would say that raising the percent black by 10 percentage points worsens test scores by 1/2th of a standard deviation. (A 10 percentage point change in percent black typically translates into a change in the race of two students. This is a change that is big enough to be interesting.) We should not, however, interpret these results causally because 55 percent of blacks in Connecticut live in the five largest school districts. The results in column I reflect the characteristics of these five cities.

In column II, I add the demographic variables and the indicator for mixed-grade classes. These controls for observed characteristics do not notably alter the pattern of results. The point estimates are similar and, although the standard errors triple or quadruple, all the estimates are still statistically significant.

Column III includes district fixed effects, so that the identifying variation comes from within-district changes in percent black over time. Although we cannot confidently sign the bias, we might still expect it to be negative. In fact, the four estimates that are statistically significant at the 5 percent level are all negative. Similar statements could be made about the results in column IV, which includes both district fixed effects and district-specific linear time trends.

In column V, percent-black-in-class is instrumented with the residual of percent-black-in-enrollment from district-specific regressions that contain an intercept and a quartic time trend. The standard errors are very small: a change of 2/1000$^{\text{ths}}$ of a standard deviation in test scores for a 10 percentage point change in percent black would generally be statistically significant at the 5 percent level. Nevertheless, none of the estimates is statistically significantly different from zero. These results are confirmed by the results shown in column VI, where percent-black-in-class is instrumented by residuals from percent-black-in-births-in-grade. The estimates based on birth residuals naturally have higher standard errors, but the point estimates exhibit a pattern that is roughly similar to that of column V.

We may conclude that there is no intrinsic achievement effect of having additional black students present. That is, they appear not to have different achievement on average than white students who enroll in the same schools, and they appear not to affect the achievement of other students. The entire effect of greater exposure to blacks that we get in naive estimates appears to be due to characteristics correlated with higher steady state percentages of blacks.

The results should not be interpreted as saying that desegregation plans would have no effect on schools. Students who were involuntarily moved by these plans would, on average, be associated with the characteristics that cause the poor student achievement in estimates like those in column I. Instead, the results can be interpreted as saying that parents who observe that their child happens to be in a cohort with an unusually high percentage of blacks should not expect their child's achievement to be affected.

## VI. The Effects of Gender Composition

The natural population variation strategy can also identify the effects of gender composition. We do not expect gender composition to be strongly correlated with unobserved student, parent, and community characteristics, but parents do treat boys and girls differently and private schooling opportunities differ systematically by gender. For instance, public schools with high percentages of females may be located in communities that support all-male private schools (and thus have the characteristics associated with such communities). Alternatively, parents might be more likely to remove their child from a school where violence or drugs are prevalent if that child is a boy. If so, public schools with high percentages of females might be poor environments. Using natural population variation is useful *not* because we think that the phenomena that cause schools' gender composition to be correlated with unobserved characteristics are prevalent, but because there is so little systematic variation in schools' gender composition that a significant share of it may be generated by phenomena that are relatively uncommon.

The most obvious policy related to gender composition is single-sex education, but this is rarely used in public education. The two most relevant policies are (i) modification of teaching techniques to take account of the ways in which gender composition affects the classroom and (ii) partial isolation of students by gender for learning certain subjects. These policies are based on the hypotheses that females get less attention when they are in class with a higher percentage of males and that co-educational classes generate social pressure for females to neglect math and science. Alternatively, one might hypothesize that females in disproportionately male classes are exposed to more math and science material and might, as a result, achieve higher math and science scores.

Vital statistics in Connecticut are not compiled by gender by town, so it is not possible to instrument for percent-female-in-class using percent-female-in-births-in-grade. Otherwise, the organization of Table IX is identical to that of Table VIII. Column I contains simple regressions that pool across districts and cohorts (with cohort fixed effects). If unobserved characteristics of schools, communities, or parents affect gender composition, then the results in this column will exhibit bias. A few of the point estimates are statistically significantly different from zero at the 5 percent level, and they are all negative. If we were to interpret the results causally, we would say that a 10 percentage point increase in percent female lowers mathematics scores in the 4th and 6th grade by 1/3rd of a standard deviation.

As demographic variables, district fixed effects, and district-specific linear time trends are added to the regression, statistically significant negative estimates disappear and are replaced by some statistically significant positive estimates. Focus, however, on column V, which uses residuals of percent-female-in-enrollment from district-specific regressions that contain an intercept and a quartic time trend. In this column, where identification is based on natural population variation that is credibly random, percent female is shown to have positive effects on writing in the 4th, 6th, and 8th grades and also in the value-added specifications. A ten percentage point increase in percent female raises writing scores by between 1.5 and 2 tenths of a standard deviation. Because district-level scores are not released by gender for confidentiality reasons, we cannot know for certain whether females are just better at writing or whether classrooms with a high percentage of females are more conducive to the learning of writing (a subject that requires an unusual degree of individualized attention). However, aggregate statistics on male and female scores suggest that the increase in writing scores is too high to be attributed to female's just being better at writing. The average female's writing score exceed that of the average male by between one-half and two-thirds of a standard deviation in grades 4, 6, and 8 (depending on the grade and cohort). But, the results in column V suggest that the different between an all-female and an all-male class would be between 1.5 and 2 standard deviations. In addition, a ten percentage point increase in percent female raises 4th grade math scores by 1/10th of a standard deviation. It is difficult to attribute this improvement to females just being better at math since the average female scores slightly below the average male on the 4th grade math test.

In summary the results of Table IX suggest that gender composition affects how a classroom functions. The fact that writing is more affected than reading or math may suggest that females particularly alter the way in which teachers distribute their time among individual students. Moreover, Table IX demonstrates how the natural population variation strategy can isolate causal effects from a morass of confounding variation *if they exist*.

## VII. Conclusions

This study demonstrates how to use natural population variation to identify variation in class size and composition that can be used to consistently estimate the effects of class size and composition on student achievement. I outline two separate methods for using natural population variation. The first method, based on isolating the credibly random component of the variation in population for a grade in a school, generates empirically useful variation in class size and racial and gender composition. The second method is based on exploiting the abrupt or discontinuous changes in class size that occur when natural population variation triggers a change in the number of classes in a grade in a school. I find that this second method can only produce results that are both consistent and precise when it is applied to *within-district* changes in class size and population.

Using random fluctuations in class size generated by natural population variation, I find that reductions in class size in the policy range of 15 to 30 students have no effect on student achievement. The estimates are sufficiently precise that, if a 1 student reduction in class size improved achievement by just 3/100$^{\text{ths}}$ to 4/100$^{\text{ths}}$ of a standard deviation, I would have found statistically significant effects in math, reading, and writing. These results are confirmed by within-district estimates that are identified solely by the discontinuities in class size that occur when a class is added to or subtracted from a grade. For the policy range of 15 to 30 students, I find no evidence that class size reductions are more efficacious in early grades, in schools that contain high concentrations of disadvantaged students, or in classes that are initially larger.

These results are far less likely to suffer from omitted variables bias and endogeneity bias than estimates that depend on common sources of variation in inputs, such as the variation between districts. I demonstrate that common methods produce results that display expected patterns of bias. I also demonstrate that discontinuity-based results that do not rely *solely* on the discontinuities generated by class size thresholds are likely to be biased.

The natural population variation I employ has 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. The difference between this paper's results and the results of policy experiments like Tennessee's Project Star are probably due to the fact that Project Star participants were aware of being evaluated and had incentives to use the opportunities to improve achievement. This interpretation of the contrast between the results suggests that policies that only provide more resources will be less efficacious than policies that tie resources to performance. In other words, giving schools more *opportunities* to improve achievement may not be enough to guarantee that achievement is actually improved.

I also examine the effect of classroom racial composition and find that the presence of black students in a class, in and of itself, has no effect on achievement. This should not be interpreted as evidence that desegregation plans, which move students to schools in which they would not otherwise not enroll, will have no effect on achievement. It should be interpreted as evidence that a parent whose child happens to be a class that has an unusually high proportion of black students relative to the district norm should not expect his child's achievement to be affected. I also demonstrate that estimates that rely on between-district comparisons (even when they control for demographics) suffer from very substantial bias due to the correlation between racial composition and unobserved student, parental, and community characteristics.

Finally, I examine the effect of classroom gender composition and show that more female classes perform significantly better in writing in the 4th through 8th grades and in math in the 4th grade. A comparison between the improvement in writing and math performance and typical male-female differences on the tests suggest that gender composition alters classroom conduct.

## References

Angrist, Joshua, and Victor Lavy, "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement," NBER Working Paper No. 5888, 1997.

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, "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100 (1992), 1-40.

----, "School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina," *Journal of Economic Perspectives*, X:4 (Fall,1996), 31-40.

Ferguson, Ronald, "Evidence that Schools can Narrow the Black-White Test Score Gap," Malcolm Weiner Center for Social Policy, John F. Kennedy School of Government Working Paper No. H-97-04, 1997.

Hanushek, Eric, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24 (1986), 1141-77.

----, "Measuring Investment in Education," *Journal of Economic Perspectives*, X:4 (Fall, 1996), 9-30.

Heckman, James, Anne Layne-Farrar, and Petra Todd, "Does Measured School Quality Really Matter? An Examination of the Earnings-Quality Relationship," in Gary Burtless, ed., *Does Money Matter? The Link between Schools, Student Achievement, and Adult Success*. (Washington, D.C.: The Brookings Institution, 1995).

Krueger, Alan, "Experimental Estimates of Education Production Functions," NBER Working Paper No. 6051, 1997.

National Center for Education Statistics, *Digest of Education Statistics*. (Washington, D.C.: Government Printing Office, 1997).

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).

1. There is, however, debate about *how much* reductions in class size improve opportunities--particularly once class size is under 25 students.

2. National Center for Education Statistics, 1997. 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 teachers who have regular classroom instruction as their major duty.

3. The literature on racial and gender composition of classrooms, both popular and academic, is too voluminous to survey. In just the past two years, hundreds of articles have been published by education journals and newspapers on the effects of altering classroom racial and gender composition. Many school districts have experimented with single-sex classes in math and science, and a few large city districts (Philadelphia, Baltimore, and New York) have created single-sex schools. Detroit, Milwaukee, and Minneapolis have created all-black school.

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

5. See Krueger [1997] for a description of the Project Star results.

6. However, it is generally not transitory from the point of view of a cohort of students.

7. 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.

8. I also estimate such an equation for each district. Below, I discuss the advantages and disadvantages of estimating the relationship between class size and enrollment at, alternatively, the school and district level.

9. A high-order polynomial is possible because each district has multiple grades, schools, and years. Even the smallest districts contain 138 observations with which to estimate equation (6), and typical districts contain between 414 and 690 observations. Alternatively, I could use other specifications, such as splines, to pick up the nonlinearities.

10. This is an application of the general principle that the first-stage equation can be reduced-form, especially when instrumental variables estimation depends on exclusion restrictions for identification.

11. It is not obvious whether equation (1) should have $ln(C)$ or $C$ as the independent variable. $ln(C)$ is the better variable if achievement varies with percentage reduction in class size--a reduction from 20 to 18 students (a 10 percent reduction) is more valuable, say, than a reduction from 30 to 28 students (a 6.7 percent reduction). On the other hand, $C$ may be the better variable if there are decreasing returns to class size reductions. I estimated results both for $ln(C)$ and $C$. Since there was no substantive difference in the results, I show the results for $C$, which is the more familiar form of the independent variable.

12. I have also estimated the probits instrumenting for the independent variables with $ln(u)$. This method ensures that the estimated probit parameters are not functions of $\epsilon$, but is probably an unnecessary precaution for the cross-section discontinuity method and certainly an unnecessary precaution for the within-district discontinuity method (because variation within a district around the fixed rules identifies the estimates). In any case, the estimates that result are not substantively different from the results shown in section VI.

13. Even a cursory look at the data reveals the importance of state dependence. Its presence is not surprising, given that collective bargaining is nearly ubiquitous among Connecticut teachers. All Connecticut union contracts contain tenure clauses, and the majority of contracts constrain how and when a district can move a teacher between grades and schools in a district.

14. That is, the instrument (the change in predicted class size generated by a rule being triggered) is uncorrelated with the actual change in class size. The results are similar if the thresholds are selected based on a slightly different probability, such as 60 percent.

15. That is, we can write the change between this year's class size and the previous year's class size as:

$$\left(\frac{E_t}{n_{t-1}+I_t^{add}-I_t^{subtract}}-\frac{E_{t-1}}{n_{t-1}}\right)+\left(\frac{E_t}{n_{t-1}}-\frac{E_t}{n_{t-1}}\right)=\left[\frac{E_t}{n_{t-1}+I_t^{add}-I_t^{subtract}}-\frac{E_t}{n_{t-1}}\right]+\left\{\frac{E_t}{n_{t-1}}-\frac{E_{t-1}}{n_{t-1}}\right\}.$$

The term in the square brackets is the part of the change in class size that is due to a change in the number of classes. (8) is the square bracketed term with $I^{add}$ and $I^{sub}$ replaced by their predicted counterparts $\hat{I}^{add}$ and $\hat{I}^{subtract}$, which are generated by the class size thresholds. The term in the braces is the part of the change in class size that is likely to reflect underlying changes in the population between year *t* and year *t-1*.

16. The first-differenced version of equation (1) has an intercept because *X* includes indicator variables for cohorts.

17. In Connecticut, a child is ordinarily enrolled in kindergarten if he will be 5 by January 1 of the school year.

18. Some districts combine two small towns. In such cases, the towns' vital statistics are aggregated.

19. Over this period, the tests have remained consistent to allow comparisons across years. I do not use the language arts score because it is partly an aggregate of the reading and writing scores and does not appear to provide much additional information.

20. Connecticut does not appear to have a state-wide maximum class size rule, although state aid for certain programs is a function of class size.

21. In a previous version of this paper, I estimated the enrollment residuals using only those years for which I also had achievement data. The results do not differ substantively.

22. Outside of the largest districts, there was no case of an aide being assigned to a regular class for routine instruction and mixed-grade classes made up less than 4 percent of all classes. Special education students sometimes have aides. Interestingly, nearly all of the instruction that did not fit the standard elementary school model occurred in the 10 largest districts in Connecticut. Because these districts have little useful natural population variation, they can be dropped from the analysis without significantly affecting the results.

23. These are not quite value-added specifications because the students in a cohort vary slightly between grades as students move into and out of the district.

24. That is, Table 5 presents specifications that are frequently requested and interesting in substance. Specifications that are only interesting as functional form tests are not shown, though several were estimated. These include regressions that use the natural log of class size, regressions that are unweighted (that is, regressions that ignore the fact that the test scores are district averages), regressions that use enrollment residuals based only on the eleven years of data for which elementary test scores are available, and regressions that use 9th grade test scores and population-by-grade data from the Enumeration of Children.

25. 27 percent of the classes have more than 23 students, and large classes are particularly likely to be caused by natural population variation. However, no district in Connecticut appears to have a target class size above 23.

26. When considering the number of usable observations, it is helpful to keep in mind that the achievement equations have to be estimated at the district level. In order to take full advantage of the school-level data on class size, enrollment, and number of classes by grade, I used the following two-stage least squares procedure. The first-stage equations (the regression of actual class size on predicted class size based on the threshold functions) were estimated at the school level, separately for each grade. The predictions from these regressions were aggregated to the district level, and district-level achievement was regressed on these district-level predictions in the second-stage. When the first-stage regressions are restricted to depend only on observations at or around the discontinuities, the first-stage predictions are weighted so that the district-level discontinuity-based change in predicted class size is correct. For instance, if a district contains two schools of equal size, and a threshold is triggered in only one school, then the district's discontinuity-based predicted change in class size is equal to half the school's discontinuity-based predicted change in class size.

27. Secondary school administrators can create classroom segregation by inducing black students to self-select into lower level classes. In elementary schools, classroom segregation can generally be achieved only by forcing it upon students (since they are too young to choose for themselves). Such unsubtle techniques for creating segregated elementary classes in a school are increasingly rare, though still occasionally documented. Thus, an elementary school principal typically has little policy control over the racial composition of classes in his schools. In Connecticut, classes within a grade within a school tend to have nearly identical racial and gender composition.

**Table I**
Structure of Data

| class of | school-age population by grade from Vital Statistics and/or Enumeration of Children | public-school enrollment by grade from Conn. Dept. of Ed. | class size, class racial & gender composition by grade from Conn. Dept. of Ed. or CPEC | Statewide Test Scores | | | |
|---|---|---|---|---|---|---|---|
| | | | | grade 4 | grade 6 | grade 8 | grade 9 |
| 1983 | Enum & Vital | ✓ | ✓ | | | | ✓ |
| 1984 | Enum & Vital | ✓ | ✓ | | | | ✓ |
| 1985 | Enum & Vital | ✓ | ✓ | | | | ✓ |
| 1986 | Enum & Vital | ✓ | ✓ | | | | ✓ |
| 1987 | Enum & Vital | ✓ | ✓ | | | | ✓ |
| 1988 | Enum & Vital | ✓ | ✓ | | | | ✓ |
| 1989 | Enum & Vital | ✓ | ✓ | | | | |
| 1990 | Enum & Vital | ✓ | ✓ | | | | |
| 1991 | Enum & Vital | ✓ | ✓ | | | ✓ | |
| 1992 | Enum & Vital | ✓ | ✓ | | | ✓ | |
| 1993 | Vital only | ✓ | ✓ | | ✓ | ✓ | |
| 1994 | Vital only | ✓ | ✓ | | ✓ | ✓ | |
| 1995 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 1996 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 1997 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 1998 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 1999 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 2000 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 2001 | Vital only | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 2002 | Vital only | ✓ | ✓ | ✓ | ✓ | | |
| 2003 | Vital only | ✓ | ✓ | ✓ | ✓ | | |
| 2004 | Vital only | ✓ | ✓ | ✓ | | | |
| 2005 | Vital only | ✓ | ✓ | ✓ | | | |

**Table II**

OLS Estimates of the Effects of Class Size on Student Test Scores in Math, Reading, and Writing[1]

Each cell below dotted line contains the estimate from a separate regression. See next page for notes.

| | | I | II | III | IV |
|---|---|---|---|---|---|
| dependent variable[2] | independent variable | cohort fixed effects | cohort fixed effects & demographic controls | school district fixed effects & cohort fixed effects[3] | school district fixed effects, district-specific linear time trends, cohort fixed effects |
| 4th grd math score | avg class size up to 4th grd | -0.168<br>(0.013) | -0.039<br>(0.007) | -0.065<br>(0.004) | 0.001<br>(0.004) |
| 4th grd reading score | avg class size up to 4th grd | -0.080<br>(0.018) | -0.002<br>(0.025) | 0.002<br>(0.010) | 0.011<br>(0.014) |
| 4th grd writing score | avg class size up to 4th grd | -0.077<br>(0.007) | -0.019<br>(0.005) | -0.100<br>(0.005) | -0.012<br>(0.004) |
| 6th grd math score | avg class size up to 6th grd | -0.200<br>(0.014) | -0.001<br>(0.009) | -0.015<br>(0.002) | 0.001<br>(0.003) |
| 6th grd reading score | avg class size up to 6th grd | -0.185<br>(0.012) | -0.012<br>(0.007) | -0.018<br>(0.001) | -0.011<br>(0.003) |
| 6th grd writing score | avg class size up to 6th grd | -0.075<br>(0.006) | -0.001<br>(0.006) | -0.073<br>(0.003) | -0.050<br>(0.005) |
| 8th grd math score | avg class size up to 6th grd | -0.166<br>(0.017) | -0.014<br>(0.009) | -0.014<br>(0.002) | 0.019<br>(0.007) |
| 8th grd reading score | avg class size up to 6th grd | -0.142<br>(0.014) | -0.017<br>(0.008) | -0.018<br>(0.002) | 0.021<br>(0.008) |
| 8th grd writing score | avg class size up to 6th grd | -0.079<br>(0.009) | -0.015<br>(0.008) | -0.044<br>(0.003) | 0.017<br>(0.012) |
| diff betw 6th & 4th grd math scores | avg class size in 5th and 6th grds | -0.028<br>(0.003) | -0.007<br>(0.003) | 0.013<br>(0.002) | -0.009<br>(0.002) |
| diff betw 6th & 4th grd reading scores | avg class size in 5th and 6th grds | -0.044<br>(0.004) | -0.003<br>(0.003) | -0.003<br>(0.001) | 0.010<br>(0.001) |
| diff betw 6th & 4th grd writing scores | avg class size in 5th and 6th grds | -0.001<br>(0.002) | -0.002<br>(0.002) | -0.016<br>(0.002) | -0.022<br>(0.004) |
| diff betw 8th & 6th grd math scores | avg class size in 6th grd | -0.001<br>(0.004) | -0.006<br>(0.004) | 0.015<br>(0.004) | 0.015<br>(0.006) |
| diff betw 8th & 6th grd reading scores | avg class size in 6th grd | -0.003<br>(0.004) | -0.003<br>(0.004) | -0.007<br>(0.005) | -0.007<br>(0.007) |
| diff betw 8th & 6th grd writing scores | avg class size in 6th grd | -0.008<br>(0.005) | -0.002<br>(0.005) | 0.003<br>(0.001) | 0.006<br>(0.002) |

[1] All regressions are weighted by number of students over whom the dependent variable is averaged; include a fixed effect for each cohort; and have 1793 observations. See Table 1 for the structure of the panel data. Standard errors in parentheses.

[2] The dependent variables are formed by dividing the average test score by the overall standard deviation of scores on that test in Connecticut. Thus, the coefficients show how test scores, *measured in standard deviations*, change when class size changes by one student.

[3] The controls are: median household income, percentage of the population in poverty, percentage of adults who are college graduates, percentage of the population who are black, percentage of the population who are Hispanic, and an indicator for the class containing two grades. Only the indicator for two-grade classes remains in the equation when district fixed effects are included (columns III and IV).

**Table III**

Explanatory Power of Instruments: F-statistics on Joint Significance of Instruments in First-Stage Regressions[1]

Each cell below dotted line contains the estimate from a separate regression.

| | I | II | III | IV | V |
|---|---|---|---|---|---|
| | independent variables are residuals of... | | | | |
| | enrollment-in-grade... | | | | births-in-grade... |
| | from grade-district-specific regressions on an intercept & a... | | | | |
| | linear<br>time trend | quadratic<br>time trend | cubic<br>time trend | quartic<br>time trend | quartic<br>time trend |
| 1st grd class size | 13.060<br>(0.000) | 15.000<br>(0.000) | 9.400<br>(0.000) | 11.180<br>(0.000) | 5.872<br>(0.000) |
| 2nd grd class size | 13.748<br>(0.000) | 12.082<br>(0.000) | 12.768<br>(0.000) | 10.080<br>(0.000) | 4.292<br>(0.000) |
| 3rd grd class size | 12.870<br>(0.000) | 12.600<br>(0.000) | 11.280<br>(0.000) | 14.960<br>(0.000) | 4.171<br>(0.000) |
| 4th grd class size | 12.680<br>(0.000) | 14.480<br>(0.000) | 13.740<br>(0.000) | 10.580<br>(0.000) | 3.950<br>(0.000) |
| 5th grd class size | 13.260<br>(0.000) | 12.480<br>(0.000) | 11.340<br>(0.000) | 15.720<br>(0.000) | 3.762<br>(0.000) |
| 6th grd class size | 12.507<br>(0.000) | 14.278<br>(0.000) | 12.507<br>(0.000) | 12.353<br>(0.000) | 2.814<br>(0.000) |

[1] There are 1793 observations in each regression. Probability>F in parentheses. Each first stage regression has, as its dependent variable, class size in a grade. Each regression has, as its independent variables: the 1st through the 8th powers of an estimate of *ln(u)*. These are the excluded instruments to which the title of the table refers. Each regression also contains cohort fixed effects, district fixed effects, and district-specific linear time trends. Not surprisingly (because of the residuals' construction), the inclusion of the fixed effects and time trends has a negligible effect on the estimated effects of the excluded instrument. (The fixed effects and time trends are included simply to facilitate comparison of estimates across specifications.) Similar but slightly larger F-statistics are obtained if the dependent variable is the natural log of class size in grade.

**Table IV**

IV Estimates of the Effects of Class Size on Student Test Scores in Math, Reading, and Writing[1]

Each cell below dotted line contains the estimate from a separate regression. See next page for notes.

| | | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| | | independent variable is instrumented with the residuals of... | | | | |
| | | enrollment-in-grade... | | | | births-in-grade... |
| | | from grade-district-specific regressions on an intercept & a... | | | | |
| dependent variable[2] | independent variable | linear time trend | quadratic time trend | cubic time trend | quartic time trend | quartic time trend |
| 4th grd math score | avg class size up to 4th grd | -0.014<br>(0.017) | -0.008<br>(0.017) | -0.013<br>(0.016) | -0.010<br>(0.017) | -0.021<br>(0.090) |
| 4th grd reading score | avg class size up to 4th grd | -0.009<br>(0.006) | -0.008<br>(0.006) | -0.006<br>(0.006) | -0.005<br>(0.006) | 0.019<br>(0.036) |
| 4th grd writing score | avg class size up to 4th grd | 0.026<br>(0.014) | 0.026<br>(0.014) | 0.023<br>(0.014) | 0.023<br>(0.014) | -0.035<br>(0.073) |
| 6th grd math score | avg class size up to 6th grd | -0.004<br>(0.015) | 0.002<br>(0.014) | 0.010<br>(0.016) | 0.007<br>(0.016) | 0.047<br>(0.076) |
| 6th grd reading score | avg class size up to 6th grd | -0.008<br>(0.016) | -0.004<br>(0.016) | -0.013<br>(0.017) | -0.011<br>(0.017) | -0.007<br>(0.085) |
| 6th grd writing score | avg class size up to 6th grd | 0.030<br>(0.018) | 0.031<br>(0.017) | 0.030<br>(0.020) | 0.015<br>(0.019) | 0.005<br>(0.089) |
| 8th grd math score | avg class size up to 6th grd | 0.010<br>(0.008) | -0.011<br>(0.014) | 0.006<br>(0.009) | 0.001<br>(0.010) | 0.002<br>(0.057) |
| 8th grd reading score | avg class size up to 6th grd | 0.017<br>(0.014) | -0.001<br>(0.010) | 0.010<br>(0.008) | 0.015<br>(0.013) | -0.012<br>(0.078) |
| 8th grd writing score | avg class size up to 6th grd | 0.020<br>(0.018) | -0.012<br>(0.015) | 0.029<br>(0.017) | -0.001<br>(0.013) | -0.012<br>(0.080) |
| diff betw 6th & 4th grd math scores | avg class size in 5th and 6th grds | 0.007<br>(0.017) | 0.013<br>(0.017) | 0.016<br>(0.018) | 0.020<br>(0.018) | -0.006<br>(0.036) |
| diff betw 6th & 4th grd reading scores | avg class size in 5th and 6th grds | -0.018<br>(0.014) | -0.014<br>(0.014) | -0.011<br>(0.014) | -0.004<br>(0.014) | 0.010<br>(0.047) |
| diff betw 6th & 4th grd writing scores | avg class size in 5th and 6th grds | 0.015<br>(0.019) | 0.027<br>(0.020) | 0.030<br>(0.019) | 0.021<br>(0.019) | -0.011<br>(0.031) |
| diff betw 8th & 6th grd math scores | avg class size in 6th grd | 0.016<br>(0.010) | 0.018<br>(0.010) | 0.020<br>(0.011) | 0.012<br>(0.012) | -0.012<br>(0.031) |
| diff betw 8th & 6th grd reading scores | avg class size in 6th grd | -0.015<br>(0.011) | -0.013<br>(0.011) | -0.015<br>(0.011) | -0.015<br>(0.012) | 0.013<br>(0.027) |

| | | | | | | |
|---|---|---|---|---|---|---|
| diff betw 8th & 6th grd writing scores | avg class size in 6th grd | 0.019<br>(0.019) | 0.013<br>(0.018) | 0.010<br>(0.019) | 0.008<br>(0.021) | 0.001<br>(0.059) |

[1] Standard errors are in parentheses. All regressions are weighted by number of students over whom the dependent variable is averaged, and all regressions have 1793 observations. The specification is the same as column IV of Table 2, except that class size in instrumented as described at the top of the table. See Table 1 for the structure of the panel data. See Table 3 for the explanatory power of the instruments.

[2] The dependent variables are formed by dividing the average test score by the overall standard deviation of scores on that test in Connecticut. Thus, the coefficients show how test scores, *measured in standard deviations*, change when class size changes by one student.

**Table V**

IV Estimates of the Effects of Class Size, Generated by Some Alternative Specifications[1]

Each cell below dotted line contains the estimate from a separate regression. See next page for notes.

| | I | II | III | IV | V | VI | VII |
|---|---|---|---|---|---|---|---|
| | independent variable, which is instrumented by residuals of enrollment-in-grade from grade-district-specific regressions on an intercept & a quartic time trend, is... | | | | | | |
| dependent variable[2] | class size most recent grd[3] | avg class size in grades 1-3 | avg class size in grades 1-2 | avg class size in grade 1 | class size spline at 23[4] | indicator for class size ≥30[5] | indicator for class size ≤15[5] |
| 4th grade | -0.031 | -0.010 | 0.011 | -0.001 | 0.001 | 0.107 | 0.048 |
| math score | (0.021) | (0.017) | (0.013) | (0.024) | (0.019) | (0.125) | (0.115) |
| 4th grade | -0.013 | -0.005 | -0.007 | -0.013 | -0.011 | 0.063 | -0.026 |
| reading score | (0.014) | (0.006) | (0.004) | (0.023) | (0.007) | (0.058) | (0.046) |
| 4th grade | -0.030 | 0.023 | 0.011 | 0.018 | 0.030 | 0.132 | -0.016 |
| writing score | (0.016) | (0.014) | (0.010) | (0.019) | (0.016) | (0.107) | (0.096) |
| 6th grade | 0.001 | 0.017 | 0.008 | 0.031 | 0.004 | -0.068 | 0.128 |
| math score | (0.014) | (0.017) | (0.013) | (0.026) | (0.017) | (0.147) | (0.094) |
| 6th grade | -0.001 | 0.020 | 0.020 | 0.018 | -0.016 | -0.275 | 0.158 |
| reading score | (0.014) | (0.019) | (0.013) | (0.029) | (0.018) | (0.151) | (0.092) |
| 6th grade | -0.020 | 0.030 | 0.021 | 0.019 | 0.016 | 0.026 | 0.141 |
| writing score | (0.013) | (0.024) | (0.015) | (0.015) | (0.021) | (0.152) | (0.116) |
| 8th grade | 0.012 | 0.022 | 0.030 | 0.001 | -0.011 | 0.135 | 0.244 |
| math score | (0.009) | (0.026) | (0.028) | (0.014) | (0.018) | (0.151) | (0.191) |
| 8th grade | -0.015 | 0.037 | 0.028 | 0.015 | -0.008 | -0.207 | 0.158 |
| reading score | (0.008) | (0.026) | (0.028) | (0.012) | (0.010) | (0.139) | (0.179) |
| 8th grade | 0.016 | -0.016 | -0.027 | -0.006 | -0.001 | -0.171 | 0.313 |
| writing score | (0.011) | (0.027) | (0.026) | (0.016) | (0.023) | (0.175) | (0.209) |
| diff betw 6th & | 0.015 | 0.027 | 0.017 | 0.001 | 0.022 | 0.001 | -0.025 |
| 4th grd math | (0.018) | (0.024) | (0.017) | (0.020) | (0.022) | (0.102) | (0.200) |
| diff betw 6th & | -0.002 | 0.023 | 0.019 | 0.035 | 0.015 | 0.001 | 0.133 |
| 4th grd reading | (0.013) | (0.019) | (0.013) | (0.028) | (0.018) | (0.103) | (0.183) |
| diff betw 6th & | -0.007 | 0.008 | 0.014 | 0.004 | 0.022 | -0.001 | -0.008 |
| 4th grd writing | (0.018) | (0.015) | (0.020) | (0.022) | (0.020) | (0.103) | (0.117) |
| diff betw 8th & | 0.002 | 0.008 | 0.012 | 0.010 | 0.004 | 0.009 | 0.004 |
| 6th grd math | (0.012) | (0.024) | (0.025) | (0.018) | (0.017) | (0.207) | (0.104) |
| diff betw 8th & | -0.015 | -0.023 | -0.009 | 0.051 | -0.001 | -0.209 | 0.139 |
| 6th grd reading | (0.012) | (0.022) | (0.025) | (0.028) | (0.020) | (0.187) | (0.103) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| diff betw 8th & | 0.008 | -0.001 | -0.026 | -0.005 | 0.024 | 0.101 | -0.006 |
| 6th grd writing | (0.021) | (0.017) | (0.016) | (0.017) | (0.026) | (0.186) | (0.114) |

[1] Standard errors are in parentheses. All regressions are weighted by number of students over whom the dependent variable is averaged, and all regressions have 1793 observations. The specification is the same as column IV of Table 2, except that class size in instrumented as described at the top of the table.

[2] The dependent variables are formed by dividing the average test score by the overall standard deviation of scores on that test in Connecticut.

[3] The most recent grade is 3rd grade for the 4th grade scores, 5th grade for the 6th grade scores, 6th grade for the 8th grade scores, 5th grade for the difference between the 6th and 4th grade scores, and 6th grade for the difference between the 8th and 6th grade scores.

[4] Column V shows the results of a spline specification. The estimate shown is the estimated coefficient for class size *above* 23.

[5] The independent variable in column VI is an indicator variable for class size of 30 or more having existed in at least two out of the three of the most recent grades. The independent variable in column VII is constructed similarly, but indicates class size of 15 or less.

**Table VI**

IV Estimates of the Effects of Class Size, Differentiated by Demographics of School District.[1] See next page for notes and continuation of table.

| dependent variable[2] | independent variable | each row below is a regression | | | each row below is a regression | | | each row below is a regression | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | low inc | med inc | high inc | high pov | med pov | low pov | high %blk | med %blk | low %blk |
| 4th grd math score | avg class size up to 4th grd | -0.035<br>(0.019) | -0.011<br>(0.018) | -0.017<br>(0.020) | -0.028<br>(0.019) | -0.024<br>(0.022) | -0.003<br>(0.019) | -0.007<br>(0.017) | 0.009<br>(0.021) | 0.001<br>(0.025) |
| 4th grd reading score | avg class size up to 4th grd | -0.005<br>(0.007) | 0.004<br>(0.006) | -0.001<br>(0.007) | -0.003<br>(0.007) | -0.001<br>(0.008) | 0.004<br>(0.007) | 0.001<br>(0.006) | 0.005<br>(0.007) | 0.008<br>(0.009) |
| 4th grd writing score | avg class size up to 4th grd | 0.012<br>(0.015) | 0.009<br>(0.015) | 0.022<br>(0.017) | 0.013<br>(0.016) | 0.006<br>(0.018) | 0.016<br>(0.016) | 0.010<br>(0.014) | -0.002<br>(0.017) | 0.028<br>(0.020) |
| 6th grd math score | avg class size up to 6th grd | -0.008<br>(0.014) | 0.006<br>(0.015) | -0.007<br>(0.016) | 0.003<br>(0.013) | -0.003<br>(0.013) | -0.001<br>(0.017) | 0.004<br>(0.012) | -0.009<br>(0.013) | 0.003<br>(0.016) |
| 6th grd reading score | avg class size up to 6th grd | -0.008<br>(0.016) | -0.013<br>(0.016) | 0.002<br>(0.018) | -0.016<br>(0.013) | -0.005<br>(0.013) | 0.024<br>(0.018) | -0.015<br>(0.013) | 0.002<br>(0.014) | -0.002<br>(0.018) |
| 6th grd writing score | avg class size up to 6th grd | 0.018<br>(0.017) | 0.021<br>(0.018) | -0.006<br>(0.019) | -0.003<br>(0.015) | 0.016<br>(0.015) | -0.003<br>(0.021) | 0.017<br>(0.015) | 0.019<br>(0.016) | 0.003<br>(0.020) |
| 8th grd math score | avg class size up to 6th grd | 0.021<br>(0.038) | 0.012<br>(0.011) | -0.081<br>(0.109) | 0.007<br>(0.038) | 0.013<br>(0.011) | -0.053<br>(0.080) | 0.011<br>(0.019) | 0.013<br>(0.010) | -0.009<br>(0.066) |
| 8th grd reading score | avg class size up to 6th grd | 0.018<br>(0.046) | 0.021<br>(0.013) | -0.198<br>(0.131) | 0.052<br>(0.047) | 0.020<br>(0.013) | -0.086<br>(0.097) | 0.018<br>(0.026) | 0.023<br>(0.014) | -0.092<br>(0.090) |
| 8th grd writing score | avg class size up to 6th grd | 0.048<br>(0.069) | 0.022<br>(0.020) | -0.181<br>(0.198) | 0.106<br>(0.065) | 0.017<br>(0.018) | -0.194<br>(0.134) | 0.041<br>(0.037) | 0.010<br>(0.020) | -0.108<br>(0.125) |
| diff betw 6th & 4th grd math scores | avg class size in 5th & 6th grds | 0.020<br>(0.016) | -0.014<br>(0.023) | 0.005<br>(0.029) | 0.017<br>(0.015) | -0.018<br>(0.020) | -0.018<br>(0.023) | 0.025<br>(0.016) | -0.015<br>(0.013) | 0.016<br>(0.017) |
| diff betw 6th & 4th grd reading scores | avg class size in 5th & 6th grds | 0.007<br>(0.012) | -0.019<br>(0.017) | -0.009<br>(0.022) | -0.008<br>(0.012) | -0.031<br>(0.016) | -0.030<br>(0.018) | -0.007<br>(0.011) | -0.028<br>(0.014) | -0.027<br>(0.016) |
| diff betw 6th & 4th grd writing scores | avg class size in 5th & 6th grds | 0.025<br>(0.015) | 0.020<br>(0.019) | 0.018<br>(0.027) | 0.015<br>(0.015) | -0.001<br>(0.019) | 0.009<br>(0.021) | 0.014<br>(0.013) | -0.001<br>(0.018) | 0.008<br>(0.019) |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| diff betw 8th & 6th | avg class size | 0.017 | 0.056 | 0.014 | 0.007 | 0.023 | 0.082 | 0.007 | -0.002 | -0.129 |
| grd math scores | in 6th grd | (0.011) | (0.030) | (0.060) | (0.012) | (0.039) | (0.074) | (0.012) | (0.035) | (0.089) |
| diff betw 8th & 6th | avg class size | -0.016 | 0.049 | 0.023 | -0.016 | 0.009 | 0.062 | -0.015 | 0.017 | 0.035 |
| grd reading scores | in 6th grd | (0.012) | (0.032) | (0.064) | (0.013) | (0.041) | (0.078) | (0.013) | (0.036) | (0.094) |
| diff betw 8th & 6th | avg class size | 0.011 | 0.036 | 0.074 | 0.013 | 0.067 | 0.049 | 0.013 | 0.001 | -0.011 |
| grd writing scores | in 6th grd | (0.018) | (0.044) | (0.074) | (0.020) | (0.038) | (0.065) | (0.019) | (0.035) | (0.053) |

[1] Each regression has the same specification as that of column IV in Table 4, except that the class size variable is interacted with indicator variables for a demographic characteristic of the school district. Each set of 3 indicator variables is mutually exclusive. For instance, the indicator for a low income district is equal to 1 if the district has median household income less than or equal to the 25th 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 25th percentile and less than 75th 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 75th percentile; 0 otherwise. The indicators for low, medium, and high poverty are constructed similarly around the 25th and 75th percentiles of the poverty rate in Connecticut districts. The indicators for low, medium, and high percent black are constructed similarly around the 25th and 75th percentiles of the percentage of the population that is black in Connecticut districts.

**Table VII**

IV Estimates of the Effects of Class Size, Generated by Regression Discontinuity Specifications[1]

Each cell below dotted line contains the estimate from a separate regression. See next page for notes and continuation of table.

| | I | II | III | IV | V |
|---|---|---|---|---|---|
| | cross-section method[3] | | | | within-district method[4] |
| | using the predicted class size function | | | | |
| | | | – | | |
| dependent variable[2] | in its entirety | within 8 students of a discontinuity | within 4 students of a discontinuity | solely at the discontinuities | solely at the discontinuities |
| 4th grade math score | -0.053<br>(0.018)<br>[-2.944] | 0.003<br>(0.027)<br>[0.110] | -0.001<br>(0.036)<br>[-0.031] | 0.039<br>(0.089)<br>[0.441] | -0.019<br>(0.017)<br>[-1.118] |
| 4th grade reading score | -0.062<br>(0.026)<br>[-2.432] | -0.034<br>(0.038)<br>[-0.901] | -0.017<br>(0.051)<br>[-0.340] | -0.116<br>(0.126)<br>[-0.916] | 0.001<br>(0.007)<br>[0.079] |
| 4th grade writing score | -0.008<br>(0.018)<br>[-0.459] | 0.051<br>(0.027)<br>[1.900] | 0.041<br>(0.036)<br>[1.132] | 0.021<br>(0.091)<br>[0.229] | -0.029<br>(0.019)<br>[-1.526] |
| 6th grade math score | -0.126<br>(0.017)<br>[-7.511] | -0.088<br>(0.022)<br>[-3.995] | -0.063<br>(0.029)<br>[-2.142] | -0.046<br>(0.066)<br>[-0.689] | 0.001<br>(0.011)<br>[0.082] |
| 6th grade reading score | -0.134<br>(0.016)<br>[-8.288] | -0.096<br>(0.021)<br>[-4.511] | -0.067<br>(0.029)<br>[-2.348] | 0.032<br>(0.033)<br>[0.966] | 0.011<br>(0.012)<br>[0.936] |
| 6th grade writing score | -0.057<br>(0.021)<br>[-2.766] | -0.028<br>(0.027)<br>[-1.049] | 0.017<br>(0.037)<br>[0.456] | -0.019<br>(0.085)<br>[-0.224] | -0.002<br>(0.027)<br>[-0.091] |
| 8th grade math score | -0.056<br>(0.013)<br>[-4.313] | -0.054<br>(0.017)<br>[-3.223] | -0.050<br>(0.022)<br>[-2.263] | 0.022<br>(0.050)<br>[0.432] | 0.009<br>(0.015)<br>[0.648] |
| 8th grade reading score | -0.093<br>(0.012)<br>[-7.915] | -0.070<br>(0.015)<br>[-4.569] | -0.046<br>(0.021)<br>[-2.220] | -0.056<br>(0.047)<br>[-1.191] | 0.012<br>(0.011)<br>[1.053] |

| | | | | | |
|---|---|---|---|---|---|
| 8th grade writing | -0.012 | 0.000 | 0.025 | 0.093 | -0.012 |
| score | (0.016) | (0.020) | (0.028) | (0.064) | (0.043) |
| | [-0.760] | [-0.009] | [0.915] | [1.452] | [-0.286] |
| diff betw 6th & 4th | -0.060 | -0.038 | -0.038 | -0.048 | 0.018 |
| grd math | (0.011) | (0.013) | (0.019) | (0.047) | (0.023) |
| | [-5.666] | [-2.872] | [-2.001] | [-1.029] | [0.783] |
| diff betw 6th & 4th | -0.056 | -0.033 | -0.027 | -0.006 | -0.021 |
| grd reading | (0.010) | (0.013) | (0.019) | (0.046) | (0.025) |
| | [-5.379] | [-2.510] | [-1.427] | [-0.124] | [-0.820] |
| diff betw 6th & 4th | -0.053 | -0.038 | -0.032 | 0.009 | 0.067 |
| grd writing | (0.013) | (0.016) | (0.022) | (0.058) | (0.054) |
| | [-4.117] | [-2.370] | [-1.435] | [0.161] | [1.237] |
| diff betw 8th & 6th | -0.013 | -0.010 | 0.001 | -0.027 | -0.011 |
| grd math | (0.006) | (0.007) | (0.010) | (0.025) | (0.030) |
| | [-2.212] | [-1.390] | [0.074] | [-1.106] | [-0.382] |
| diff betw 8th & 6th | -0.006 | -0.001 | 0.005 | 0.022 | -0.040 |
| grd reading | (0.006) | (0.007) | (0.011) | (0.026) | (0.034) |
| | [-0.936] | [-0.123] | [0.512] | [0.858] | [-1.170] |
| diff betw 8th & 6th | -0.016 | -0.024 | -0.016 | -0.020 | 0.418 |
| grd writing | (0.010) | (0.012) | (0.018) | (0.042) | (0.301) |
| | [-1.661] | [-1.921] | [-0.900] | [-0.487] | [1.389] |

[1] All regressions are weighted by the typical number of observations over which the dependent variable is averaged. Standard errors in parentheses; t-statistics in square brackets. The independent variable is class size in most recent grade, instrumented by predicted class size. There are 1793 observations in column I, and the t-statistics on predicted class size in the first-stage regressions are 14.8 (rows 1-3), 14.1 (rows 4-6 and 10-12), and 6.7 (rows 7-9 and 13-15). The numbers of effective observations in column II are 1019 (rows 1-3), 1100 (rows 4-6 and 10-12), and 1017 (rows 7-9and 13-15), and the t-statistics on predicted class size in the first-stage regressions are 10.1 (rows 1-3), 9.4 (rows 4-6 and 10-12), and 4.6 (rows 7-9 and 13-15). The corresponding numbers in column III are 535, 559, 525, 7.9, 7.5, and 3.8. The corresponding numbers in column IV are 127, 104, 122, 2.2, 2.7, and 2.0. The corresponding numbers in column V are 567, 652, 792, 8.4, 13.4, and 3.8.

[2] The dependent variables are formed by dividing the average test score by the overall standard deviation of scores on that test in Connecticut.

[3] 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 the each district's maximum class size and the formula given by equation (7). See text for further explanation.

[4] The within-district method exploits the fact the Connecticut data are panel data and actual changes in the number of classes are observed. The equation is estimated in first-differences: the change in scores between back-to-back cohorts regressed on the change in class size, instrumented by the change in the predicted class size function. The only changes in the predicted class size function used are those generated by a change in the number of classes.

**Table VIII**

OLS and IV Estimates of the Effects of Racial Composition on Student Test Scores in Math, Reading, and Writing[1]

Each cell below dotted line contains the estimated coefficient from a separate regression. See next page for notes and continuation of table.

| | | I | II | III | IV | V | VI |
|---|---|---|---|---|---|---|---|
| dependent variable[2] | independent variable | cohort fixed effects | cohort fixed effects & demographic controls | school district fixed effects & cohort fixed effects | school district & cohort fixed effects, district-specific linear time trends | same as previous column, but independent var is instrumented by enrollment residuals | same as previous column, but independent var is instrumented by birth residuals |
| 4th grd math score | %black in class up to 3rd grd | -0.049<br>(0.001)<br>[-50.820] | -0.024<br>(0.004)<br>[-6.704] | 0.003<br>(0.007)<br>[0.440] | -0.015<br>(0.007)<br>[-2.244] | -0.014<br>(0.008)<br>[-1.675] | -0.014<br>(0.034)<br>[-0.409] |
| 4th grd reading score | %black in class up to 3rd grd | -0.024<br>(0.002)<br>[-14.514] | -0.017<br>(0.009)<br>[-1.952] | -0.009<br>(0.018)<br>[-0.497] | -0.026<br>(0.024)<br>[-1.092] | -0.028<br>(0.026)<br>[-1.068] | -0.022<br>(0.127)<br>[-0.173] |
| 4th grd writing score | %black in class up to 3rd grd | -0.025<br>(0.001)<br>[-42.432] | -0.013<br>(0.002)<br>[-5.871] | 0.008<br>(0.005)<br>[1.786] | -0.005<br>(0.006)<br>[-0.774] | -0.004<br>(0.006)<br>[-0.555] | -0.005<br>(0.031)<br>[-0.160] |
| 6th grd math score | %black in class up to 5th grd | -0.065<br>(0.001)<br>[-55.404] | -0.082<br>(0.004)<br>[-20.232] | -0.021<br>(0.007)<br>[-2.840] | -0.027<br>(0.009)<br>[-3.086] | -0.013<br>(0.009)<br>[-1.444] | -0.012<br>(0.044)<br>[-0.273] |
| 6th grd reading score | %black in class up to 5th grd | -0.062<br>(0.001)<br>[-57.625] | -0.074<br>(0.004)<br>[-20.653] | -0.012<br>(0.006)<br>[-1.914] | -0.025<br>(0.008)<br>[-3.299] | -0.015<br>(0.008)<br>[-1.778] | -0.017<br>(0.049)<br>[-0.345] |
| 6th grd writing score | %black in class up to 5th grd | -0.022<br>(0.001)<br>[-35.712] | -0.024<br>(0.003)<br>[-8.659] | -0.003<br>(0.006)<br>[-0.546] | 0.001<br>(0.007)<br>[0.009] | 0.005<br>(0.008)<br>[0.707] | -0.006<br>(0.041)<br>[-0.149] |
| 8th grd math score | %black in class up to 6th grd | -0.062<br>(0.001)<br>[-52.564] | -0.082<br>(0.004)<br>[-21.368] | -0.028<br>(0.006)<br>[-4.401] | -0.001<br>(0.008)<br>[-0.092] | 0.009<br>(0.008)<br>[1.073] | 0.016<br>(0.043)<br>[0.376] |
| 8th grd reading score | %black in class up to 6th grd | -0.063<br>(0.001)<br>[-56.704] | -0.061<br>(0.004)<br>[-15.828] | 0.001<br>(0.006)<br>[-0.054] | -0.008<br>(0.009)<br>[-0.929] | -0.004<br>(0.009)<br>[-0.434] | 0.001<br>(0.052)<br>[0.019] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 8th grd writing score | %black in class up to 6th grd | -0.024<br>(0.001)<br>[-32.535] | -0.035<br>(0.003)<br>[-10.311] | -0.018<br>(0.006)<br>[-2.913] | -0.011<br>(0.009)<br>[-1.297] | -0.008<br>(0.009)<br>[-0.838] | -0.013<br>(0.045)<br>[-0.287] |
| diff betw 6th & 4th grd math scores | %black in class in 4th and 5th grds | -0.022<br>(0.001)<br>[-24.562] | -0.053<br>(0.004)<br>[-13.029] | -0.002<br>(0.009)<br>[-0.273] | 0.001<br>(0.011)<br>[0.079] | 0.007<br>(0.012)<br>[0.588] | 0.007<br>(0.062)<br>[0.113] |
| diff betw 6th & 4th grd reading scores | %black in class in 4th and 5th grds | -0.041<br>(0.001)<br>[-46.728] | -0.056<br>(0.003)<br>[-17.345] | -0.009<br>(0.006)<br>[-1.526] | 0.001<br>(0.008)<br>[0.076] | 0.001<br>(0.008)<br>[0.145] | -0.002<br>(0.043)<br>[-0.047] |
| diff betw 6th & 4th grd writing scores | %black in class in 4th and 5th grds | -0.002<br>(0.001)<br>[-3.349] | -0.008<br>(0.003)<br>[-2.194] | -0.011<br>(0.008)<br>[-1.402] | 0.008<br>(0.010)<br>[0.727] | 0.017<br>(0.011)<br>[1.576] | -0.017<br>(0.056)<br>[-0.304] |
| diff betw 8th & 6th grd math scores | %black in class in 6th grd | -0.002<br>(0.001)<br>[-2.204] | -0.006<br>(0.003)<br>[-2.281] | -0.009<br>(0.007)<br>[-1.275] | 0.017<br>(0.009)<br>[1.872] | 0.018<br>(0.010)<br>[1.842] | 0.019<br>(0.056)<br>[0.339] |
| diff betw 8th & 6th grd reading scores | %black in class in 6th grd | -0.002<br>(0.001)<br>[-2.959] | 0.012<br>(0.003)<br>[3.425] | 0.010<br>(0.008)<br>[1.310] | 0.015<br>(0.010)<br>[1.449] | 0.020<br>(0.011)<br>[1.892] | 0.022<br>(0.056)<br>[0.392] |
| diff betw 8th & 6th grd writing scores | %black in class in 6th grd | -0.005<br>(0.001)<br>[-6.463] | -0.012<br>(0.004)<br>[-3.121] | -0.012<br>(0.008)<br>[-1.495] | -0.014<br>(0.012)<br>[-1.211] | -0.018<br>(0.013)<br>[-1.406] | 0.001<br>(0.064)<br>[0.016] |

[1] Except for "percent-black-in-class" having replaced class size, the specifications are exactly like those of Table 2 and columns IV and V of Table 4.

**Table IX**

OLS and IV Estimates of the Effects of Gender Composition on Student Test Scores in Math, Reading, and Writing[1]

Each cell below dotted line contains the estimated coefficient from a separate regression. See next page for notes and continuation of table.

| | | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| dependent variable[2] | independent variable | cohort fixed effects | cohort fixed effects & demographic controls | school district fixed effects & cohort fixed effects | school district & cohort fixed effects, district-specific linear time trends | same as previous column, but independent var is instrumented by enrollment residuals |
| 4th grd math score | %female in class up to 3rd grd | -0.034<br>(0.016)<br>[-2.060] | -0.004<br>(0.008)<br>[-0.515] | 0.006<br>(0.005)<br>[1.204] | 0.007<br>(0.005)<br>[1.288] | 0.010<br>(0.005)<br>[1.930] |
| 4th grd reading score | %female in class up to 3rd grd | -0.007<br>(0.021)<br>[-0.352] | 0.005<br>(0.027)<br>[0.175] | -0.007<br>(0.023)<br>[-0.294] | -0.006<br>(0.029)<br>[-0.205] | -0.014<br>(0.045)<br>[-0.307] |
| 4th grd writing score | %female in class up to 3rd grd | -0.007<br>(0.008)<br>[-0.882] | 0.007<br>(0.005)<br>[1.307] | 0.009<br>(0.004)<br>[2.493] | 0.013<br>(0.004)<br>[2.838] | 0.014<br>(0.006)<br>[2.457] |
| 6th grd math score | %female in class up to 5th grd | -0.032<br>(0.017)<br>[-1.955] | -0.008<br>(0.009)<br>[-0.894] | -0.003<br>(0.005)<br>[-0.487] | 0.008<br>(0.005)<br>[1.661] | 0.003<br>(0.007)<br>[0.461] |
| 6th grd reading score | %female in class up to 5th grd | -0.016<br>(0.016)<br>[-1.031] | 0.001<br>(0.007)<br>[0.087] | -0.001<br>(0.003)<br>[-0.183] | 0.001<br>(0.004)<br>[0.382] | -0.003<br>(0.004)<br>[-0.739] |
| 6th grd writing score | %female in class up to 5th grd | -0.003<br>(0.007)<br>[-0.409] | 0.004<br>(0.006)<br>[0.750] | 0.009<br>(0.004)<br>[2.497] | 0.014<br>(0.004)<br>[3.719] | 0.018<br>(0.005)<br>[3.581] |
| 8th grd math score | %female in class up to 6th grd | -0.024<br>(0.021)<br>[-1.123] | 0.001<br>(0.009)<br>[-0.018] | -0.002<br>(0.004)<br>[-0.430] | 0.010<br>(0.004)<br>[2.360] | 0.006<br>(0.006)<br>[1.172] |
| 8th grd reading score | %female in class up to 6th grd | -0.019<br>(0.018)<br>[-1.059] | -0.001<br>(0.008)<br>[-0.176] | -0.001<br>(0.004)<br>[-0.165] | 0.004<br>(0.005)<br>[0.812] | 0.008<br>(0.006)<br>[1.399] |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8th grd writing score | %female in class up to 6th grd | 0.007<br>(0.011)<br>[0.634] | 0.016<br>(0.008)<br>[2.088] | 0.015<br>(0.006)<br>[2.641] | 0.018<br>(0.006)<br>[2.882] | 0.021<br>(0.007)<br>[3.038] |
| diff betw 6th & 4th grd math scores | %female in class in 4th and 5th grds | -0.003<br>(0.008)<br>[-0.405] | 0.004<br>(0.008)<br>[0.520] | 0.003<br>(0.007)<br>[0.388] | 0.002<br>(0.006)<br>[0.369] | -0.002<br>(0.009)<br>[-0.270] |
| diff betw 6th & 4th grd reading scores | %female in class in 4th and 5th grds | -0.019<br>(0.010)<br>[-1.968] | 0.005<br>(0.006)<br>[0.895] | -0.002<br>(0.003)<br>[-0.585] | -0.002<br>(0.003)<br>[-0.490] | -0.003<br>(0.004)<br>[-0.869] |
| diff betw 6th & 4th grd writing scores | %female in class in 4th and 5th grds | 0.005<br>(0.005)<br>[1.065] | 0.009<br>(0.006)<br>[1.516] | 0.006<br>(0.005)<br>[1.244] | 0.005<br>(0.005)<br>[0.919] | 0.005<br>(0.003)<br>[2.081] |
| diff betw 8th & 6th grd math scores | %female in class in 6th grd | -0.003<br>(0.006)<br>[-0.535] | 0.001<br>(0.006)<br>[0.057] | -0.005<br>(0.005)<br>[-1.061] | 0.001<br>(0.005)<br>[0.274] | -0.001<br>(0.005)<br>[-0.247] |
| diff betw 8th & 6th grd reading scores | %female in class in 6th grd | 0.001<br>(0.005)<br>[0.174] | -0.003<br>(0.005)<br>[-0.584] | 0.002<br>(0.005)<br>[0.400] | 0.006<br>(0.006)<br>[1.043] | 0.004<br>(0.005)<br>[0.873] |
| diff betw 8th & 6th grd writing scores | %female in class in 6th grd | 0.001<br>(0.007)<br>[0.181] | 0.005<br>(0.008)<br>[0.649] | 0.007<br>(0.004)<br>[1.871] | 0.004<br>(0.009)<br>[0.432] | 0.004<br>(0.003)<br>[1.628] |

[1] Except for "percent-female-in-class" having replaced class size, the specifications are exactly like those of Table 2 and column IV of Table 4.

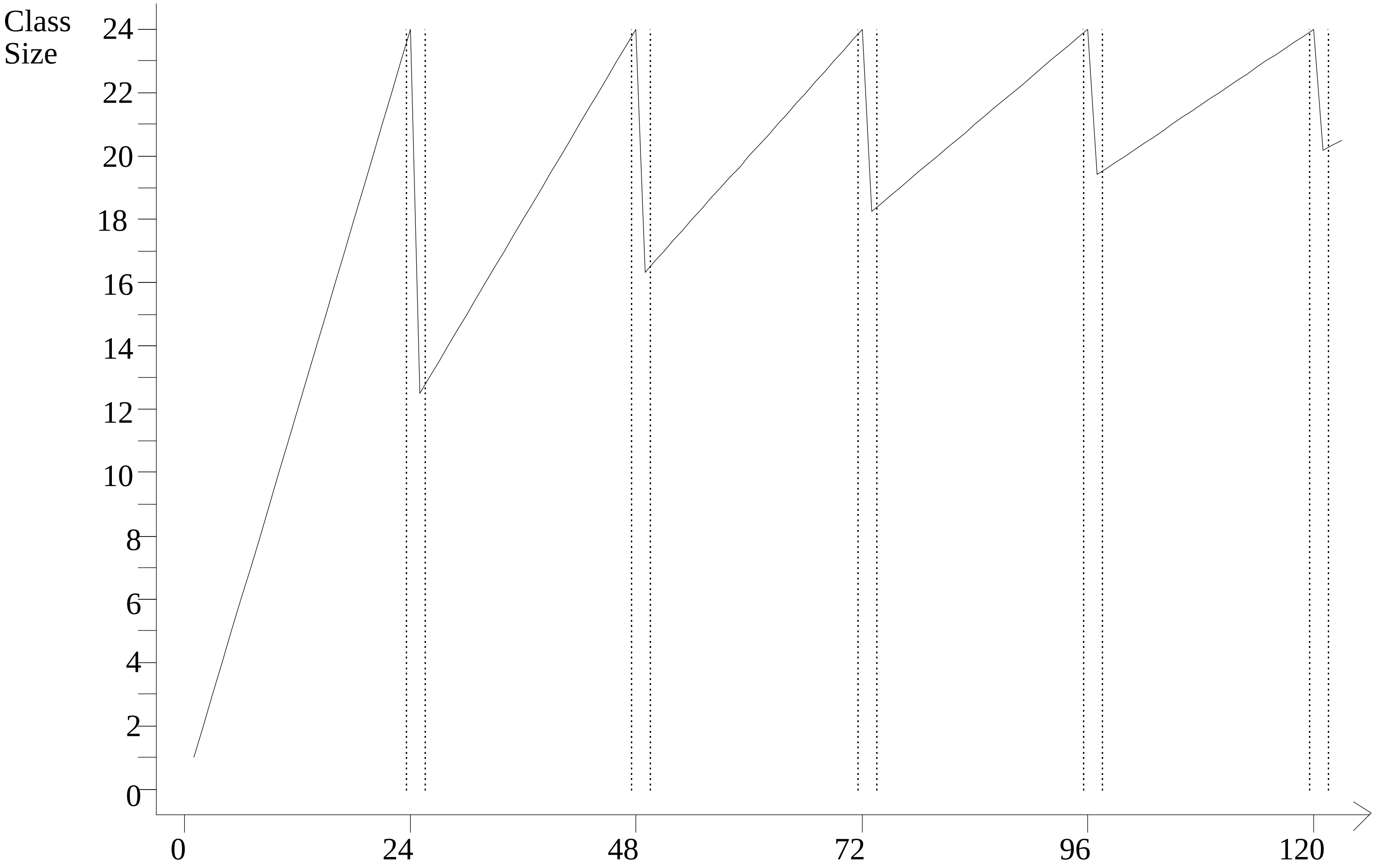
Class
Size
24
22
20
18
16
14
12
10
8
6
4
2
0
0
24
48
72
96
120
Enrollment

**Figure I**

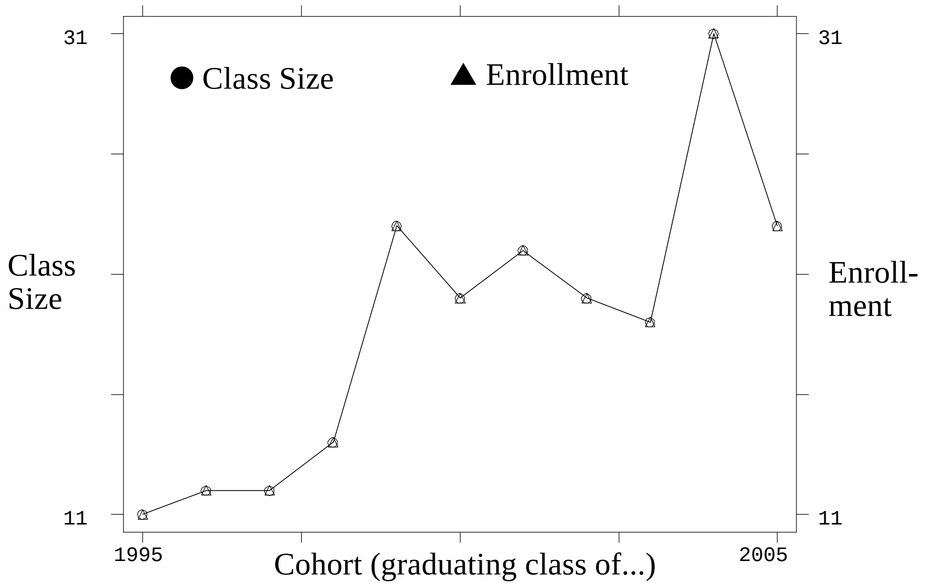

**Figure IIa**

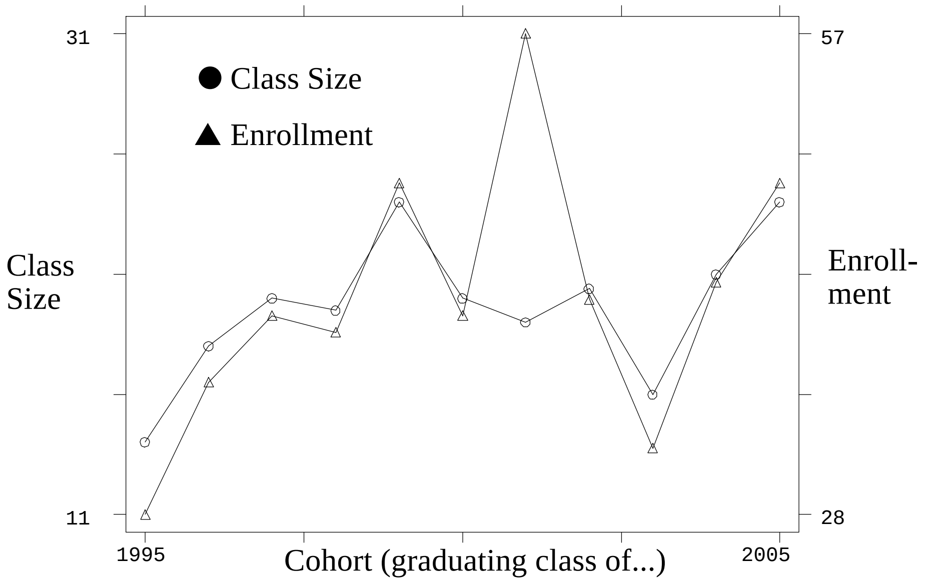

**Figure IIb**

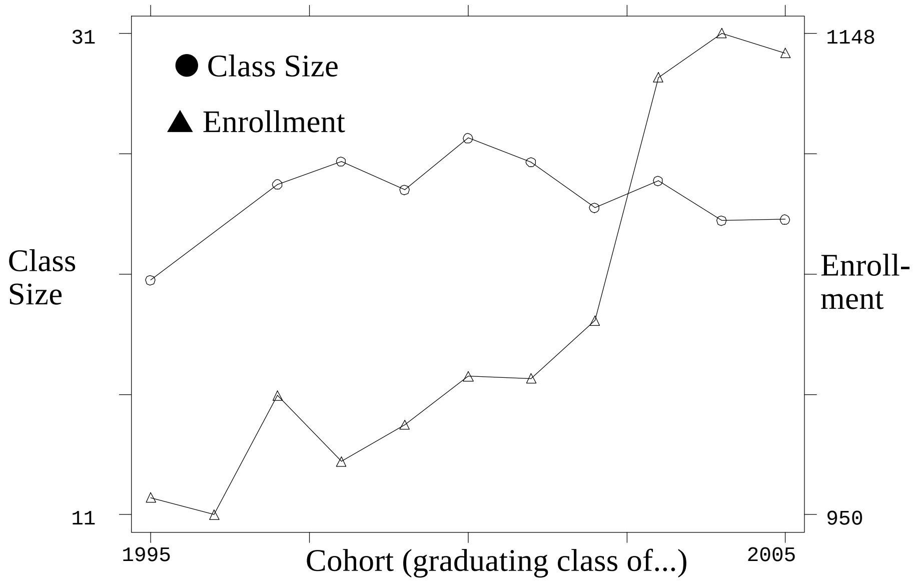

**Figure IIc**

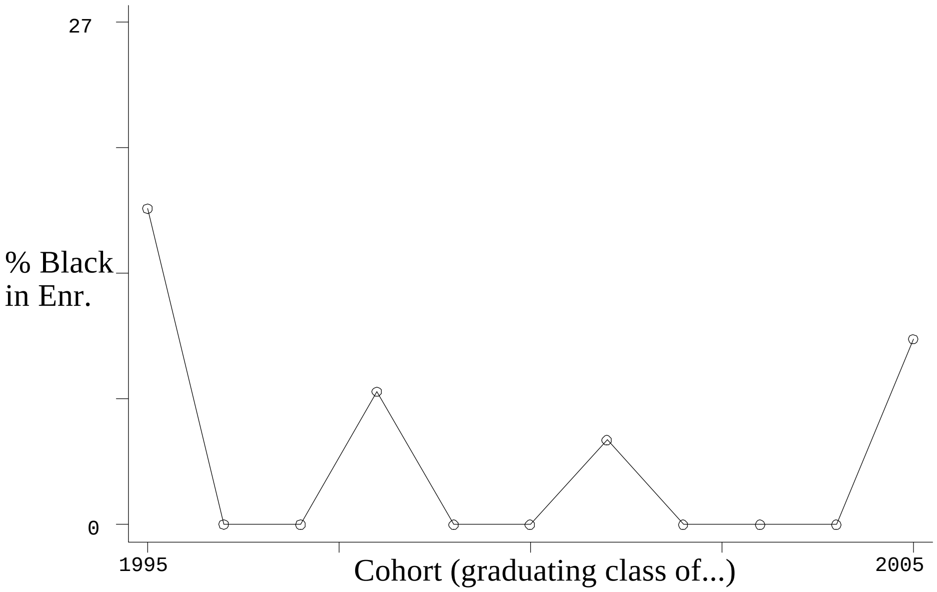

**Figure IIIa**

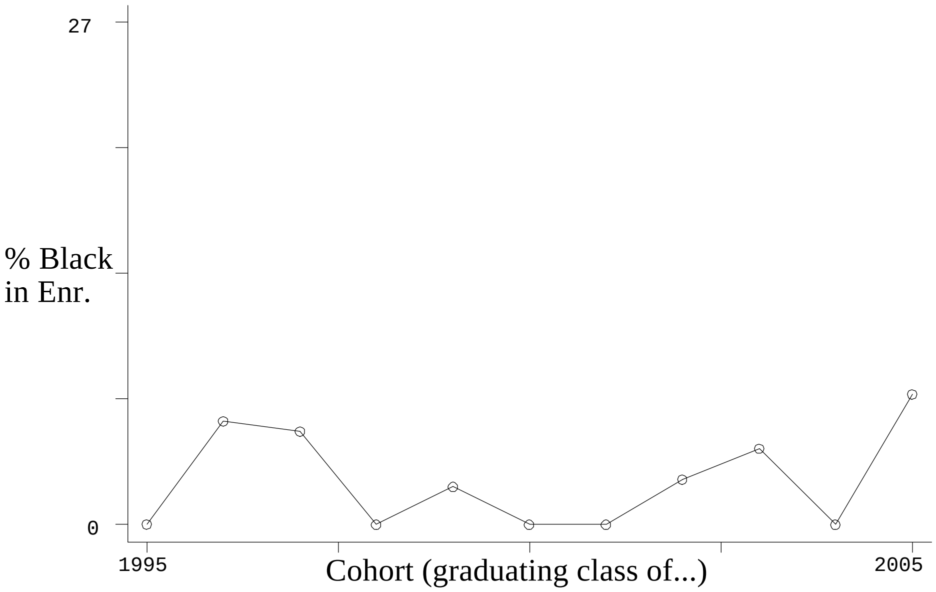

**Figure IIIb**

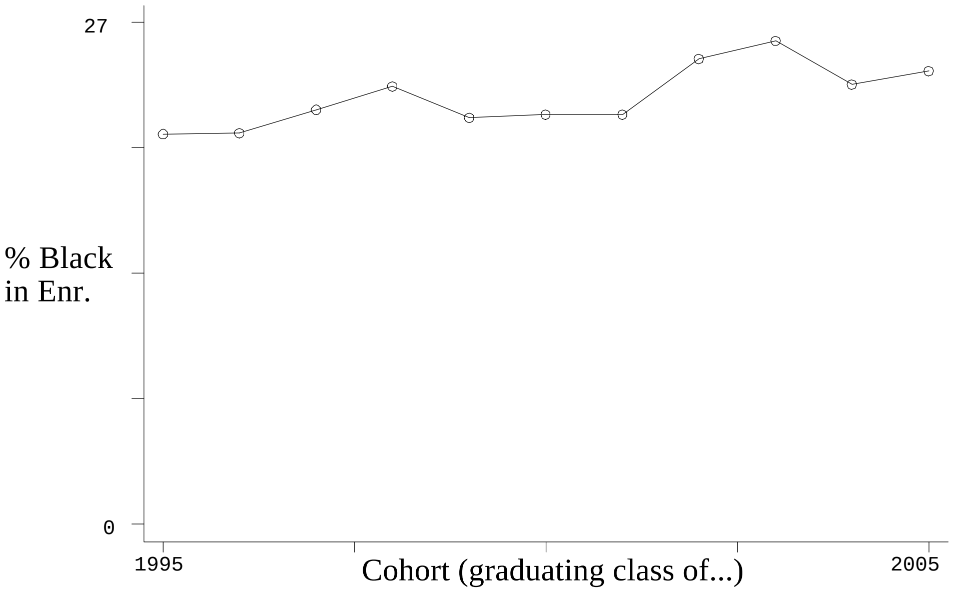

**Figure IIIc**

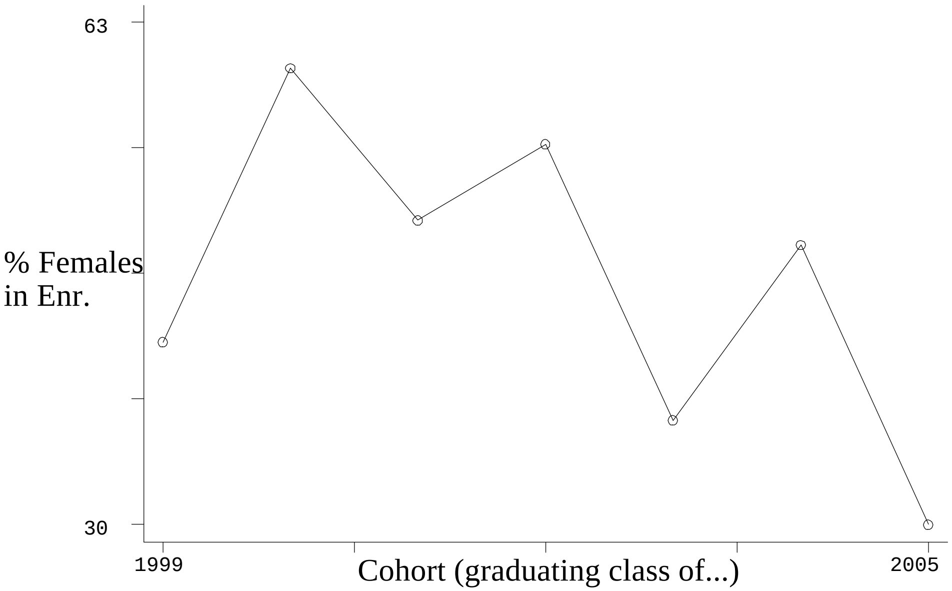

**Figure IVa**

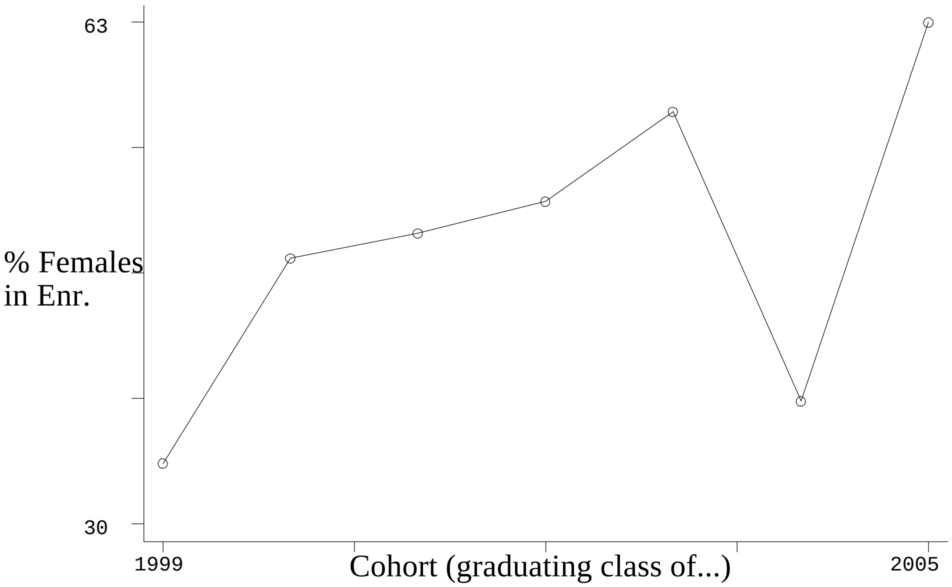

**Figure IVb**

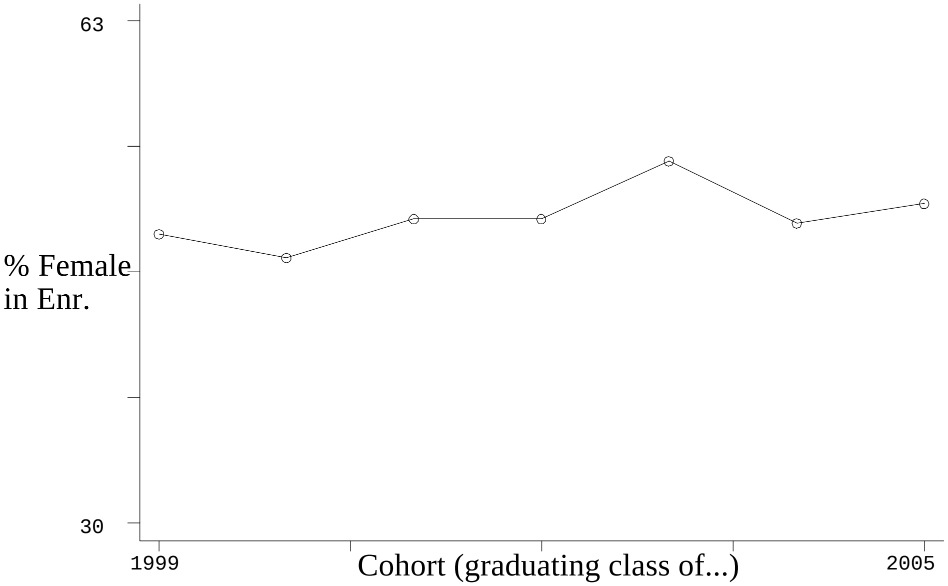

**Figure IVc**

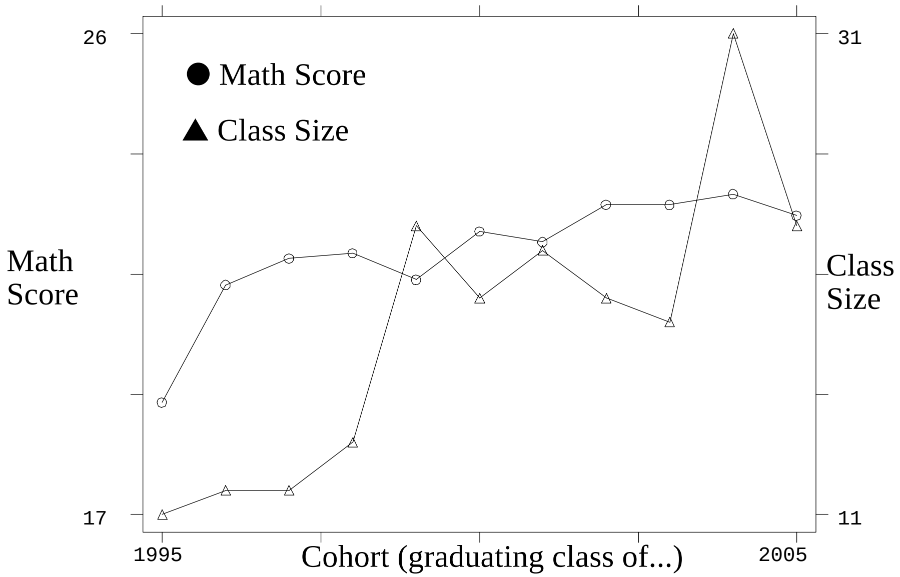

**Figure Va**

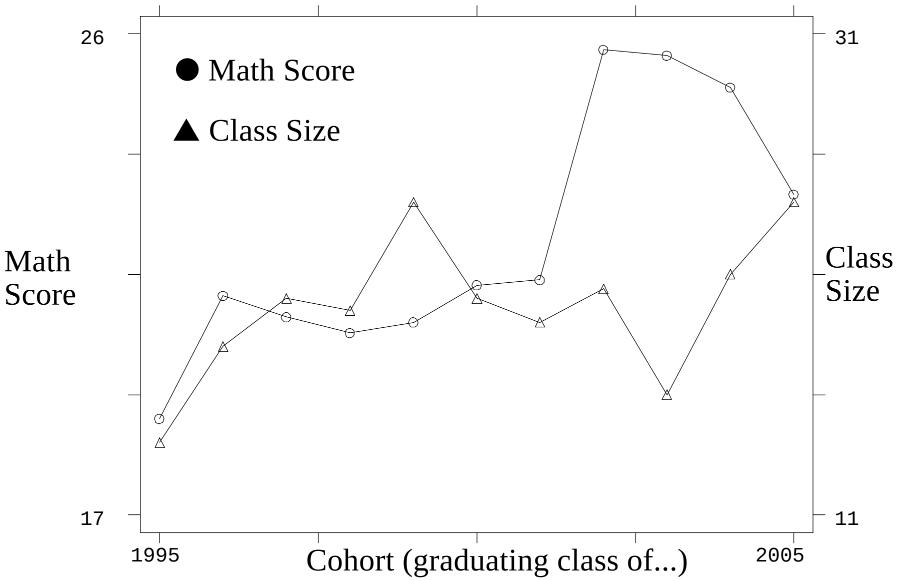

**Figure Vb**

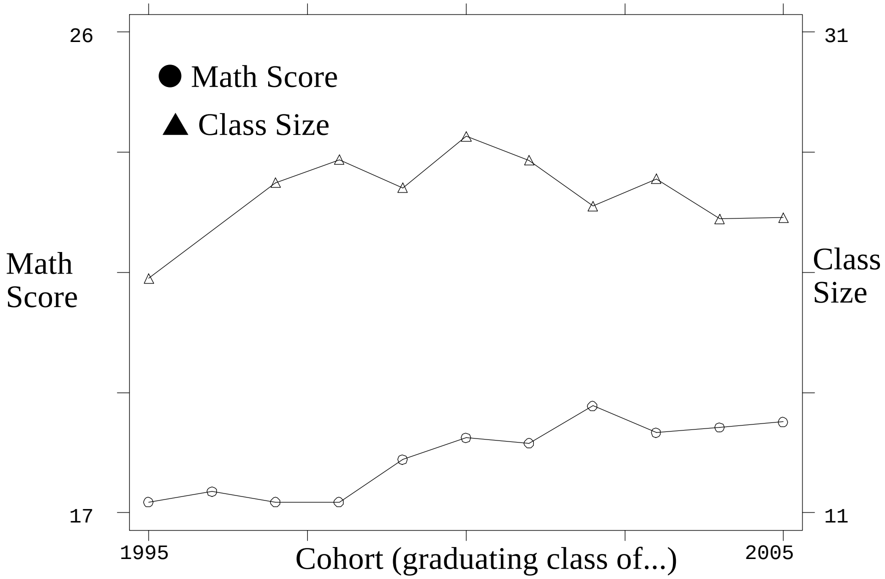

**Figure Vc**

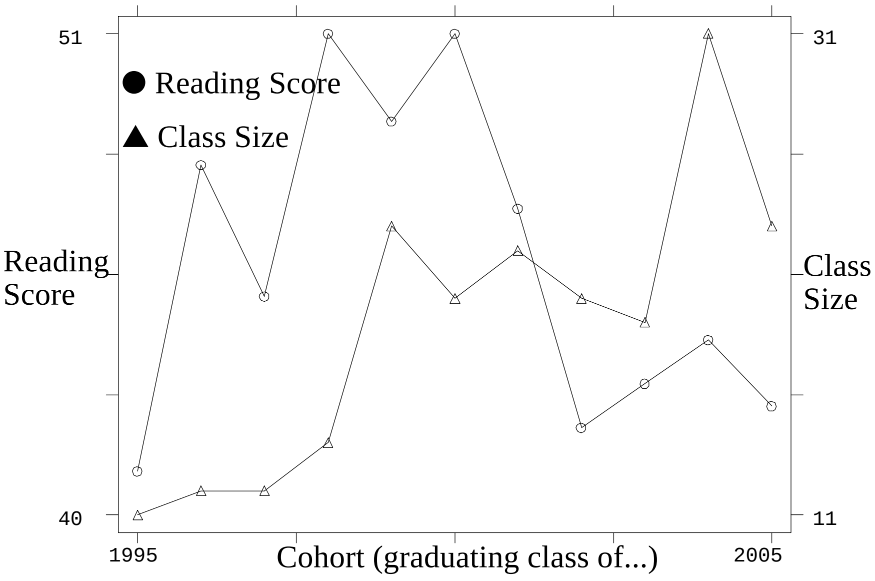

**Figure VIa**

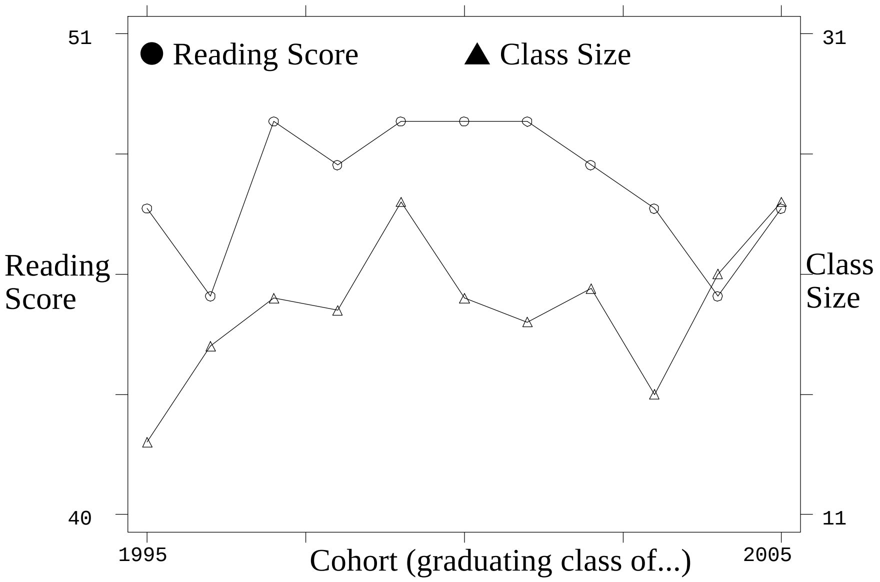

**Figure VIb**

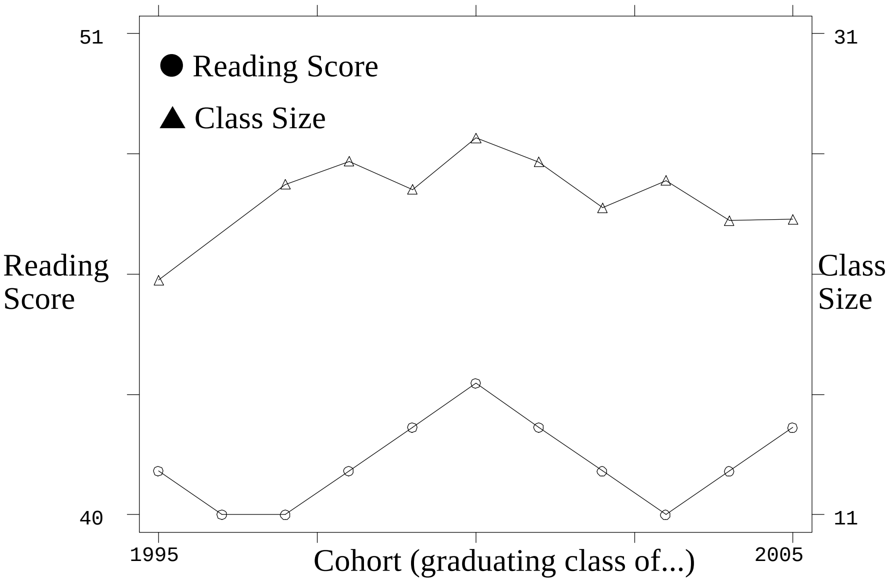

**Figure VIc**

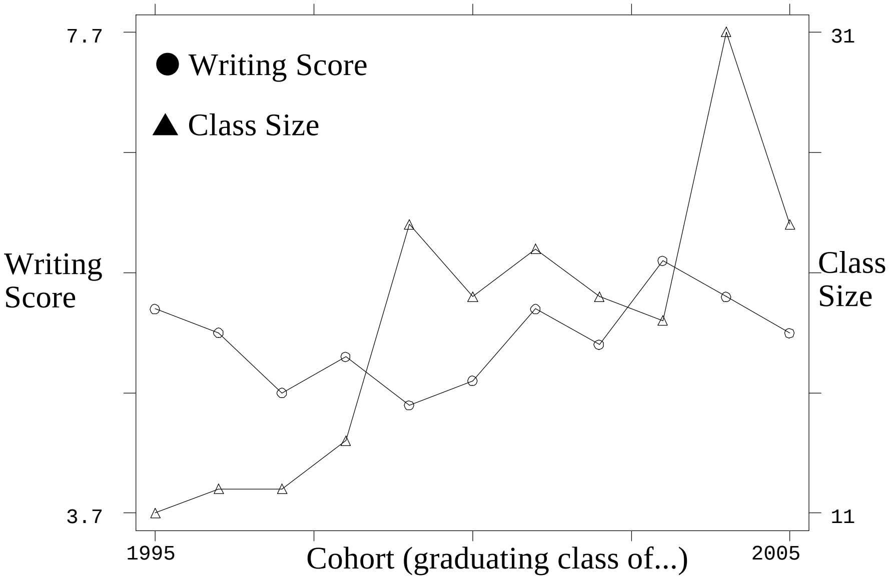

**Figure VIIa**

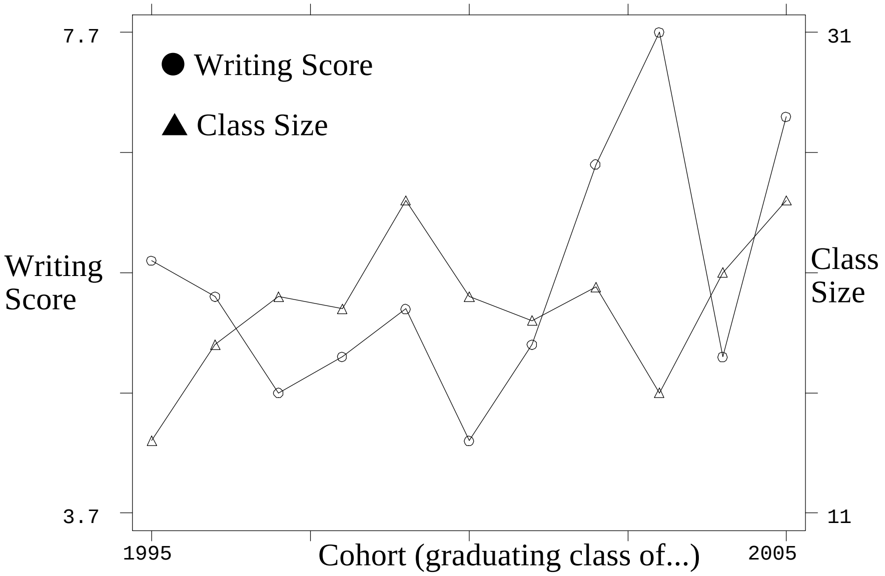

**Figure VIIb**

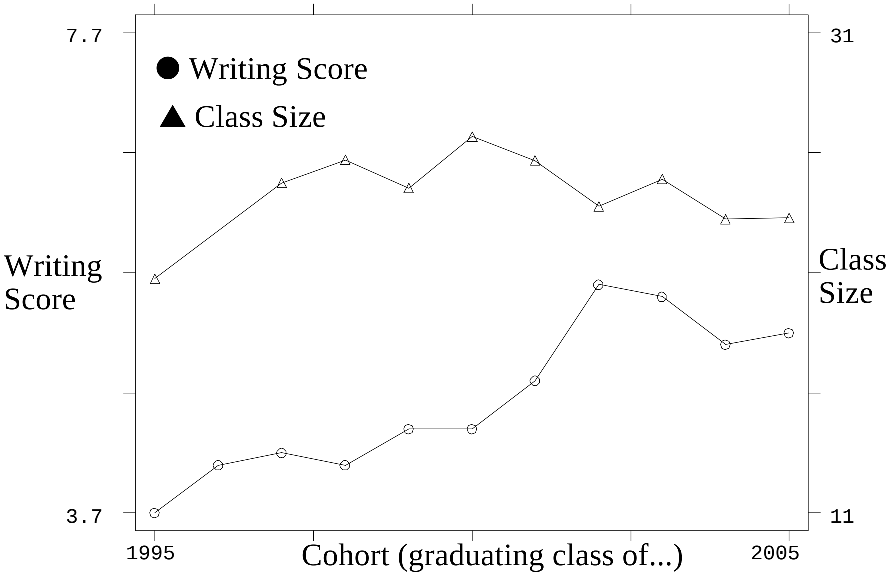

**Figure VIIc**