Revised Draft
February 6, 2001

THE CLASS SIZE CONTROVERSY

by

Ronald G. Ehrenberg, Dominic J. Brewer, Adam Gamoran and

J. Douglas Willms*

(Forthcoming in *Psychological Science in the Public Interest*)

*** Ehrenberg is the Irving M. Ives Professor of Industrial and Labor Relations and Economics at Cornell University and Director of the Cornell Higher Education Research Institute. Brewer is Director of RAND Education. Gamoran is Professor of Sociology and Educational Policy Studies at the University of Wisconsin-Madison. Willms is Professor of Education at the University of New Brunswick and Director of the Canadian Research Institute for Social Policy.** **We are grateful to four anonymous referees for their comments on an earlier draft. The views expressed herein are solely our own.**

Revision February 6, 2001

I. Introduction

A. The Learning Process

Schooling has multiple purposes. In the long run, higher levels of schooling are associated with higher earnings and economic mobility, better health, lower mortality rates, and greater democratic participation. For these reasons, most societies require children to attend school for a specified number of years or until they reach a certain age. Many of the benefits of schooling occur in part because students learn some new knowledge or skills that enhance their ability to communicate, solve problems and make decisions. Much of debate over schooling is essentially about how to maximize the amount of student learning, typically as measured by various assessment instruments such as standardized achievement tests. From a societal viewpoint, since resources – most notably time are required for learning, and are scarce, the amount of learning needs to be maximized at least cost.

Learning is complex, involving cognitive processes that are not completely understood. Typically, school systems have established a primary mode of learning that involves groups of students of about the same age interacting with a single individual leading activities in a confined physical space, directed towards learning a particular topic – in other words, students are placed in classes. The number of other students in the class can vary. At the extreme, there can be one or more adults facilitating learning – teachers -- with one or two students. At the other, a student may be one of a few hundred being taught by a single instructor (or with new Internet technology, one of millions).

The number of students in a class has the potential to affect how much is learned in a number of different ways. For example, it could affect how students interact with each other – the level of social engagement. This may result, for example, in more or less noise and disruptive behavior, which in turn affect the kinds of activities the teacher is able to promote. It could affect how much time the teacher is able to focus on individual students and their specific needs rather than on the group as a whole. Since it is easier to focus on one individual in a smaller group, the smaller the class size the more likely individual attention can be given, in theory at least. The class size could also affect the teacher's allocation of time, and hence effectiveness, in other ways too – for example, how much material can be covered. Teachers may choose different methods of teaching and assessment when they have smaller classes. For example, they may assign more writing, or provide more feedback on students' written work, or use open-ended assessments, or encourage more discussions, all activities that may be more feasible with a smaller number of students. Exposure to a particular learning environment may affect learning over during the time period of exposure – or it may have longer term or delayed effects (e.g., by increasing self-esteem or cognitive developments that have lasting effects).

For these reasons, changes to the class size are considered a potential means of changing how much students learn. Not only is class size potentially one of the key variables in the "production" of learning or knowledge, it is one of the simplest variables for policymakers to manipulate. However, the amount of student learning is dependent on many different factors. Some are related to the classroom and school environment in

which the class takes place, but others are related to the student's own background and motivation, and broader community influences.

When we ask whether class size matters for achievement, it is essential to ask also, *how* class size matters. This is important for three reasons. First, if we can observe not only achievement differences, but also the mechanisms through which the differences are produced, this will increase our confidence that the differences are real, and not an artifact of some unmeasured or inadequately controlled condition. Second, the effects of class size may vary in different circumstances, and identifying *how* class size affects achievement will help us to understand why the effects of class size are variable. Third, the *potential* benefits of class size reduction may be greater than what we observe. For example, suppose class size reductions aid achievement, but only when teachers modify instructional practices to take advantage of the smaller classes. If a few teachers make such modifications, but most do not, then understanding *how* class size affects achievement in some cases will help reveal its potential effects, even if the potential is generally unrealized.

B. The Meaning and Measurement of Class Size

Class size is not the same thing as the pupil-teacher ratio. Indeed it is quite different. The calculation of a pupil-teacher ratio typically includes teachers who spend all or part of their day as administrators, librarians, special education support staff, itinerant teachers, or other roles outside the classroom. Thus, pupil-teacher ratio is a global measure of the human resources brought to bear, directly and indirectly, on children's learning. Class size refers to the actual number of pupils taught by a teacher at

a particular time. Thus, the pupil-teacher ratio is always lower than the average class size, and the discrepancy between the two can vary, depending on teachers' roles and the amount of time teachers spend in the classroom during the school day. From an administrative or economic viewpoint, pupil-teacher ratio is very important, because it is closely related to the amount of money spent per child. However, from a psychological viewpoint – in terms of how students learn – what matters is the number of students that are physically present interacting among themselves and with the teacher. This paper focuses mainly on class size, because it is a more direct measure of the teaching resources brought to bear on a child's development.

The measurement of class size is not as straightforward as it might seem. It can vary considerably for a single child at different times during a school day and school year, because of student mobility, student absences, truancy, or the presence of pull-out special education classes. Thus, a class with 20 registered pupils will vary in its class size from day to day, and may have far fewer than 20 pupils at particular times. In the middle and secondary school grades, class size tends to vary by subject area, and therefore can vary for each pupil during a school day. Ideally, one would like to have a measure of the actual class size experienced by every pupil during every school day, over the school year.

While class size data may be available to researchers who intensively study a small number of classrooms, in practice, data on pupil-teacher ratios are more readily available to most researchers than detailed data on class sizes. Pupil-teacher ratio data can be used to examine the relationship between schooling outcomes and pupil-teacher ratios, but this relationship is likely to be weaker than the relationship between schooling

outcomes and class size, as class size is more closely linked with learning. Class size data that include a temporal dimension are seldom available; in most cases, researchers use data pertaining to the number of pupils enrolled in a class. "Class size" measures thus typically contain considerable measurement error. If this measurement error is random, estimates of the relationship between schooling outcomes and class size will be biased towards zero. That is, the relationships that are estimated will, on average, be smaller in absolute value than the true relationships between class size and school outcomes.

C. The Policy Context

The nation has struggled with how to improve its public schools for generations. Since the late 1970s and early 1980s those concerns have risen, prompted largely by threats to the nation's economic dominance and prosperity. Trends on national achievement tests have been broadly stagnant since they were begun in 1970, and international comparisons of student performance generally indicate that U.S. children, particularly in the upper grades, do not fare well. A crescendo of demands for 'reform' reached their height with the publication of *A Nation At Risk* in 1983. Since that time, many different kinds of reform have been tried – ranging from student testing and assessment; accountability systems for teachers, students and schools; new school financing arrangements; changes in the curriculum; new whole school reform designs; magnet, charter and voucher schools, etc. It has proven difficult to engender lasting or widespread change – sustaining and scaling up reform has proved extremely difficult. Translating seemingly sensible (and sometimes research-based) schemes into classroom level change breaks down in implementation. Many believe that public schools can in fact not be reformed and have turned to solutions that seek to alter the fundamental

structure of the system, such as vouchers that allow parents to choose among a range of public and private schools.

School administrators, teachers and parents have long thought that the number of children in a classroom affects the learning that occurs; however, it has proven difficult to pin down the precise effects of class size on student achievement. Various key dimensions need to be addressed from a policy perspective:

- Do students experiencing smaller class sizes learn more, as measured by student achievement tests, than otherwise similar students?
- What is the nature of the relationship between class size and student achievement – is the relationship linear, or do class sizes have to be below a certain level for a large impact to occur?
- Does the impact of class size vary across grade levels (e.g., early grades versus high school versus college), subjects being taught, school contexts (e.g., within a large school, within an urban area), and kinds of students (e.g., gifted children, at-risk students)?
- How does the timing and exposure to smaller classes affect the effects? For example, is it necessary to have several years of small classes, or just one? Does it matter whether a large class follows having a small class, or vice versa?
- Do the effects have long run consequences? Do they persist even after the treatment has ended? What things can be done differently in classes of different sizes that are not currently done?
- Why do class sizes affect (or not affect) student achievement? What is done differently, if anything, in small and large classes?

- Do the benefits of smaller class sizes outweigh the costs associated with the resources required (extra teachers, extra facilities)?
- How important is class size relative to other factors including individual student background and the mix of students, school climate, teacher behavior and quality, the nature of physical space occupied, and other resources available in the classroom?

Although we do not know the answers to these questions with any certainty, in recent years more than 20 states and the federal government have adopted various policies designed to decrease class size. Policies can vary on a number of dimensions including (Brewer, Krop, Gill & Reichart, 1999; Parish, & Brewer, 1998):

- Actual class size – some states have set the target level at 15, some at 17, others at 20
- Measurement of class size – how class size is calculated and whether the target is a class, grade, school, district or state level target
- Grade levels and subjects covered – most states have adopted policies designed to reduce class sizes in the early grades, but some have also expanded to other grades (e.g., California in grade 8) or confined the smaller classes to specific subjects (e.g., literacy)
- Kinds of students targeted. Some states have put all children in reduced size classes; others have targeted the policy to at-risk students.
- Timing – some states have phased in the policy over a number of years, while others have implemented immediately
- Funding – some states fully fund the cost of additional resources needed to create the new classes while others do not, necessitating reallocations of resources away from other programs

- Whether class size reduction is mandatory or expressed as a goal?

Thus, not only is the concept of class size somewhat difficult to pin down, the policies that are termed "class size reduction" can be very different from each other.[1] It is likely that the way a policy is designed and implemented could have major repercussions for the observed effects on student achievement.

II. Quasi-Experimental Studies Using United States Data

A. What Can We Learn from National Time Series Data

National time-series data on average class sizes faced by students are not regularly collected, but data on average pupil/ teacher ratios are collected on an annual basis. Between 1969 and 1997, the average pupil/teacher ratio nationally in American public and private elementary and secondary schools declined from 22.7 to 16.6, a decline of over 26 percent. The comparable changes in elementary and secondary schools' average pupil/teacher ratios were respectively, 25.1 to 18.3 and 19.7 to 14.0 (U.S. Department of Education, 1999). If one restricts one's attention to public schools, in which the vast majority of American children are educated, one finds very similar numbers, both in terms of levels and of changes. With such a large decline in pupil/teacher ratios, the public might have expected to observe substantial increases in the amount that students learned over the period.

Aggregate time-series evidence of changes in how much students have learned in the United States can be obtained from the *National Assessment of Educational Progress (NAEP)* tests. *NAEP* is the only national representative assessment of student knowledge in subject matter areas, and it is been conducted periodically in a number of areas,

including reading, mathematics and science.[2] A summary of the trends in average scale scores on these tests for students of different ages or grade levels is found in Figure 1. As the figure indicates, comparable data for the science test is available from 1969, for the mathematics test from 1973, and for the reading test back from 1971.

What immediately jumps out from this figure is the lack of a large increase in student test scores on any of the tests that is commensurate with the substantial decline in the average pupil/teacher ratio nationwide that took place during the period. For students at age 17, the average science test score declined between 1969 and 1984 and then increased slightly thereafter. However, by 1999 it was still lower than its 1969 level. The mathematics average score for 17 year olds declined between 1973 and 1982 then increased through 1992 and has remained roughly constant since. Finally, the reading scores of the 17 year olds rose only slightly during the period.

We have focused on the test scores of the oldest enrolled students in each year because their educational outcomes occur near graduation from high school. It is reasonable to ask, however, what the trends have been for younger students in lower grades. Perusal of Figure 1 suggests that the average mathematics test scores of 9 and 13 year olds did increase somewhat over the 1978 to 1996 period. However, even on this test, the increase in scores was not at all proportionate to the decline in the pupil/teacher ratio that had taken place. One's first inclination might be to suggest that the decline in pupil/teacher ratios did not have a major impact on students' academic achievement during the last quarter of the 20th century and to conclude that expending resources to reduce average class sizes is not a prudent investment.

There are many reasons why this inference should not be drawn from the above evidence. One is that test scores are but one measure of performance of schools. Another important measure is the years of schooling that students complete. Reducing the dropout rate is a goal of most school systems and the percentage of 16 to 24 year olds that were high school dropouts declined from 15 percent in 1970 to 11 percent in 1997 (Digest of Education Statistics, 1999). However, to the extent that potential dropouts tend to come from the lower tail of the achievement distribution and dropouts usually occur only after students reach the compulsory schooling age (16 or 17 in most states), a reduction in the dropout rate would, other factors held constant, tend to lower the average test scores of enrolled students in the upper grades. This might partially explain why the improvements in performance that took place over time on the *NAEP* tests tended to be less for 17-year-old students than for younger students.

Similarly, the percentage of individuals who graduated from high school in the preceding year who were enrolled in college rose from 51.8 to 67.0 during the same period (U.S. Bureau of the Census, 1999). While college enrollment rates are known to be sensitive to family income levels and federal and state financial aid policies changed dramatically during the period, some of the change in college enrollment rates may have been attributable to smaller pupil/teacher ratios.[3]

A second reason is that the backgrounds of students attending schools also changed in dramatic fashion during the period. For example, in 1970, 85 percent of America's students came from families with two parents in the home. By 1995, this had fallen to 68 percent (U.S. Bureau of the Census, 1999). Moreover, the probability that a married woman with children between the ages of 6 to 17 was in the labor force rose

from 49 to 76 percent during the same period (U.S. Bureau of the Census, 1999). The fraction of children who had difficulty speaking English also rose from 2.8 percent in 1979 to 5.1 percent in 1995 (U.S. Bureau of the Census, 1999).

Finally, the fraction of children living in poverty increased as the distribution of income in the United States grew more unequal. In 1970, 14.9 percent of all children under the age of 18 lived in families whose incomes fell below the poverty line, but by 1995, this percentage had risen to 20.2. The percentage of children living in families with income below the poverty line was not the same across racial/ethnic groups. While about 15.5 percent of white children came from families with incomes below the poverty line in 1995, the comparable percentages for children in African-American and Hispanic-American families were both around 40 (U.S. Bureau of the Census, 1999).

The increased incidence of children living in one-parent homes, married women with children who were in the labor force, children who had difficulty speaking English, and children living in poverty are all parts of what has been referred to as the decline in the "social capital" available to many American school children (Hedges, & Greenwald, 1996; Nye, Hedges, & Konstantopoulos, 1999). Together these factors tend to make it more difficult for some students to learn, and reduce the time and resources that their parents have to invest in their education and support their efforts. The changes in these factors might be expected to make it more difficult to tease out the effect of reductions in pupil/teacher ratios in the aggregate data.

It may well be that pupil/teacher ratio reductions matter more for some groups of students than for others. For example, it is likely that the children from lower-income families with fewer and less-educated parents living in the home are the ones that benefit

the most from smaller class sizes. In fact, there is evidence in the *NAEP* data that while African Americans and Hispanic Americans tend not to perform as well as white Americans on the *NAEP* tests, on average their performance on many of the tests increased relative to the performance of white Americans during the period that the data cover.[4]

Of course not all factors related to social capital moved in the direction of reducing parent support for children's educational achievement. The percentage of the population of young adults that had at least a high school education rose from about 74 percent in 1970 to over 85 percent in 1990. Increases in parental education should be associated with increased student learning (Hanushek, 1999).

Still a third reason is that changes in the pupil/teacher ratio were also accompanied by changes in the characteristics of the nation's teachers. For example, between 1971 and 1996, the median age of American public school teachers rose from 35 to 44, their median years of teaching experience rose from 8 to 15 years, and the fraction of teachers with at least a master's degree increased from 27.1 to 54.5 percent. Changes in the age, education and experience of teachers might be expected to influence how much students learn.

In addition, the average salary of public school teachers rose during the period. Viewed in constant 1997-98 dollars (after correcting for inflation), the average teacher was paid $37,735 in 1970-71. By 1997-98, this figure had increased to $39,385 (Digest of Education Statistics, 1999). While the increase in average teacher salaries in real terms was positive, this increase was only about 4% over the 17-year period. Moreover, average

teacher salaries actually declined relative to the salaries of other full-time workers in the economy whose education levels were comparable.

For example, during the mid 1970s, the average teacher earned about the same amount as the average female employee in the economy who had five or more years of post secondary education and was employed full-time.[5] By 1990, the average teacher was earning 7 percent less than the average female employee with five or more years of post secondary education, and the differential has widened even more since then. This is not surprising because many females who in an earlier era would have been teachers have flocked to the better paying professions that have opened up to them, including business, law and medicine.

The decline in relative teacher salaries made it more difficult to attract women into the teaching profession and appears to have been associated with a decline in the academic aptitude of the people pursuing careers in education, as measured by their performance on standardized tests such as the Graduate Record Examination (Bok, 1993). Changes in teacher academic aptitude might also be expected to be associated with how much their students learn.

Finally, starting with the passage of the *Education for All Handicapped Children Act of 1975,* there has been a substantial growth in the proportion of students classified as the special education population, which in turn has led to an even greater increase in the proportion of teachers serving this population because pupil/teacher ratios are much smaller in special education. Thus, part of the decline in the pupil/teacher ratio reflects the increased share of teacher resources going to special education students. This increased focus on special education students would have caused the aggregate

pupil/teacher ratio to decline, even if the class size experienced by students in regular classes had not declined at all. In fact, it has been estimated that perhaps one-third of the decline in the pupil/teacher ratio that took place during the 1980s was due to the growth of special education classes during that decade (Hanushek, & Rivkin, 1997).

The bottom line of this discussion is that many factors other than the pupil/teacher ratio influence how much students learn and the effectiveness of schools. To adequately control for these other factors requires one either to conduct a true experiment in which students are randomly assigned to different class sizes or to conduct an analysis of nonexperimental data in which factors other than class size are adequately taken account of in the analysis. Recent research has indeed focused on true experiments, the most notable being one that was conducted in the state of Tennessee. We will discuss the experimental research on class size effects in a later section of our paper.

Most research on class size effects, however, has used nonexperimental data. To adequately control for factor other than class size that might be expected to influence the outcomes of an educational system, one needs to have an underlying conceptual framework. It is to sketching out such a framework that we now turn.

B. A Simple Conceptual Model

What are the factors that influence the educational attainment of students at a point in time? The characteristics of all the schools that the students have attended up until that point in time should matter. These characteristics include not only the class sizes in which they have been enrolled, but also the capacities of their teachers including academic aptitude, experience, subject matter knowledge, ability to motivate students to

learn, their instructional methods and coverage of content, and their classroom management skills. They also include the physical condition of the school facilities, the educational technology that is available and is used in the schools, the number and quality of support and professional staff other than teachers and the quality of the school administrators. Put another way, a students educational attainment at time t, A(t), depends upon the whole time path of school related resources that the student has been exposed to up until that time, R(t).

However, children are in their schools for less than half of the days of a year and for only about 6 hours a day. So what goes on outside of the school may be equally, or more important, than what goes on inside of a school, in determining students' educational attainment. One important set of factors pertain to the family, such as the value that they place on education and the time and financial resources that they have been able to devote to supporting their children's schooling up until that time, F(t). On average, factors such as the family's income, the number of parents or other adults in the home, the adults' education levels, the number of the siblings and whether both parents are working are likely to influence parents' views of the importance of education and the resources that they devote to encouraging education and monitoring their children's progress.

A third set of factors relates to the characteristics of the school community in which students have been educated each year up until age t and the characteristics of the classrooms in which they were enrolled each year, S(t). Attitudes of one's peers towards education and the effort that one's peers put in may influence an individual's effort and learning. The socio demographic characteristics of one's classmates, including their race

and ethnicity, may matter. Some people assert that the matching of teachers and students by race, gender and ethnicity may also matter.[6] So too, may the way that students are grouped in classrooms. Students in classes that are heterogeneous, in terms of the "ability levels" of the students, may learn more, or less, than students enrolled in classes in which students are fairly homogeneous in terms of their "ability levels".[7] The prevailing attitudes towards academic achievement that are found in the school that a student attends each year are also likely to matter.

A final set of factors is the characteristics of the broader community, M(t), that the student lived in each year. Did the community value and support education? Is there peer pressure outside of school not to do well? Such *neighborhood* or *community* effects are often alleged to be very important. Summarizing all of these forces in one equation,

(1) $A(t)=G(R(t), F(t), S(t), M(t), e(t))$.

Here G represents some general functional form and in empirical specifications e(t) is an error term. The error term is included in equation (1) to account for the facts that not all of the relevant variables can be observed and included in an empirical model, that innate academic ability varies across students and that there is some randomness in the realizations of the educational outcomes that are observed for any particular student. Note that the discussion above provides us with no guidance about what the appropriate functional form for equation (1) should be.

Equation (1) should make it clear to the reader how difficult it is to adequately control for all of the factors, in addition to a student's current class size, that influence a student's educational attainment at a particular time. There are two general strategies. One is to randomly assign students and teachers to classes of varying sizes, and observe

students' rate of learning in the different settings. Another strategy is to collect data on the relevant factors affecting student learning, and exercise some form of statistical control. The former strategy, randomized experiments, tend to be more "internally" valid, because they explicitly control for all of the confounding factors which might affect the observed relationship between schooling outcomes and class size. But randomized experiments tend to be very expensive, and are typically conducted with small samples that cover a limited number of grades and school settings. The second strategy, commonly referred to as quasi-experiments, tends to be more "externally" valid. Researchers typically use data from large state or national surveys, and thus the findings can be generalized across different settings and for students at different ages. This approach has more stringent data requirements: if factors such as school resources, family resources, school community characteristics, and neighborhood or community features that affect achievement are also correlated with class size, then these must be included in the analysis. We discuss below the limitations of both experimental and quasi-experimental approaches, review the findings stemming from each approach, and suggest ways that future experiments could be improved.

With either approach, the researcher wants to know the functional form of the relationship between outcomes and class size, and why smaller classes might matter. For example, is there some threshold at which decreases in class size have a dramatic effect? Small class sizes may matter more in the early elementary grades, when students are becoming socialized to school and forming work habits, than they do in the high school grades. Small class sizes may matter more for children from disadvantaged backgrounds, who do not have the same resources in the home to support their education

than students from rich families. Small classes may matter more for children who have difficulty learning and need to be motivated than they do for bright highly motivated students. Finally, small classes may matter more or less depending on the methods of instruction. For example, small classes may benefit students more when instruction relies on discussion, by allowing more students to participate and be recognized, than when lecture and seatwork are the main modes of instruction.

Moreover, small class sizes may matter more for some educational outcomes than they do for others. The educational attainment of a student can be measured by the number of years of elementary and secondary schooling that the student completes, the student's test scores for a given number of years of schooling, or the student's post secondary school educational attainment. Indeed, if one takes a very market-oriented approach and argues that a major purpose of the educational system is to prepare students for the world of work, the quality of the education that a student receives may be reflected in the student's post schooling labor market earnings. Being educated in smaller classes may influence some, but not necessarily all, of these different outcome measures.

C. *Equality of Educational Opportunity: The Coleman Report* and the Birth of Educational Production Functions

The Coleman Report represented an important step in educational research (U.S. Government Printing Office, 1966). Its statistical analyses based on data collected from a representative national sample of over 570,000 students, 60,000 teachers and 4,000 principals represented the beginning of the *educational production function* literature. While the initial focus of the report was on measuring the extent of racial segregation of

American schools, the survey underlying the report collected data on third, sixth, ninth and twelfth grade students' test scores in a variety of subjects and on characteristics of the students' families, their teachers, the schools they attended and the communities in which they lived. This permitted the researchers to study whether at a point in time, students' test score levels were related to their current family characteristics, their current teachers' characteristics, the sizes of the classes in which they were currently enrolled, the characteristics of the schools in which they were currently enrolled, and the characteristics of the community in which they currently lived.

That is, the data permitted them to estimate equations of the form

(2) $A(t)= a_0+ a_1r(t)+a_2f(t) +a_3s(t) +a_4m(t) +e(t).$

In this equation a_0 is an intercept term, a_1, a_2, a_3 and a_4 are vectors of parameters and the r(t), f(t), s(t) and m(t) are vectors of current year values of variables. Put another way, they estimated models that specified that a student's test score in the grade that he or she was currently enrolled was linearly related to the values of the school resources and teacher characteristics that he or she was exposed, including class size, in that year, the characteristics of his family in that year, the characteristics of the school community at the school in which he or she was enrolled in that year, the characteristics of the community in which he or she lived in that year and a random error term.[8]

A major conclusion of the *Coleman Report* was that variations in family background characteristics and community level characteristics were much more important in explaining variations in student achievement across school than were variations in school resources, such as pupil/teacher ratios or expenditures per pupil, and variations in teacher characteristics, such as experience and degree levels. This

conclusion was reached by comparing the proportions of the variance in test scores across schools that family and community characteristics uniquely explained, relative to the proportions of the variance in test scores that school resource and teacher characteristics variables explained.

Given its importance the *Coleman Report* was subject to much scrutiny by social scientists and numerous critiques of its findings and reanalyzes of its data quickly took place.[9] Some observers pointed out that because family resources were very highly correlated with school resource and teacher characteristics variables (e.g. schools with higher income families tended to have lower pupil/teacher ratios and spend more per student) it was difficult to uniquely assign the "explanatory power of a model" to one or another set of variables.

Others pointed out that to say family and community variables "explained" most of the variation in educational outcomes across schools, conveyed no information about what the marginal impact of increasing expenditures per student or decreasing the pupil/teacher ratio would be on educational outcomes. For example, it is possible that even if variations in the pupil/teacher ratio explained little of the variation in educational outcomes across schools, that reducing average pupil/teacher ratios by a given amount could have a large impact on educational outcomes. That is, these researchers stressed that the focus of policy researchers should be on estimating the marginal impact that a proposed policy change (e.g., reductions in class sizes) would have, not on worrying about the proportion of the variance in the educational outcome that that variable can explain (Cain, & Watts, 1970).

Perhaps the most telling criticism of the *Coleman Report s* analyses is that the data that are used in it represent a snapshot taken at a point in time. Educational outcomes at time t are specified to be a function only of the current values of teacher, school, family and community variables, rather than the whole history of each set of variables that a student has experienced up until that date. Thus, there are a whole set of variables that have been omitted from the analyses that potentially can bias the results.

To counter this problem, some subsequent studies, which we will discuss below, have focused on estimating *gain score* equations. Rather than relating the level of achievement that a student has achieved at the end of a school year to the levels of school, teacher, family and community variables that the student experienced during the year, they relate the change in the student's achievement during the year to the latter variables. In doing so, they implicitly are assuming that all of the effects of prior years' school, teacher, family and community related variables are captured by the student's prior year academic achievement. Put simply, variables from prior years matter only to the extent that they determine the prior year's achievement.

This obviously is a very strong assumption and one that it is not necessarily correct. For example, suppose that we are focusing on the gain in academic achievement that third graders achieve between the end of the second and third grades and trying to relate this to the third graders' class sizes. Small class sizes in the first and second grades may influence not only the second grade achievement scores of students but also their whole trajectory of learning thereafter. To the extent that students who are in small classes in the third grade also tend to be in small classes in earlier grades, attempts to relate the growth in students' test scores between the end of the second and third grades

to their class size in the third grade, in the context of an empirical model that ignores their first and second grade class sizes, will overstate (in absolute value) the impact of third grade class size on student learning. This will occur because one will have omitted from the model variables (first and second grade class size) that are positively correlated with third grade class size and that influence the gain score in the same way that the third grade class size is postulated to influence it.

Economists refer to this as the problem of omitted variable bias. Educational psychologists who study the threats to the internal validity (the difficulty of drawing correct logical inferences) of quasi-experimental designs refer to it as the problem of the interaction of selection and maturation (Campbell, & Stanley, 1966). In the current example, students "selected" to have small class sizes in grade three also tended to have small class sizes in grades one and two, so their test scores should be expected to mature (grow) at a more rapid rate than other students' test scores, other factors held constant, even in the absence of a third grade "class size effect". While the two disciplines use different languages, the problem they refer to is the same. Excluding variables from gain score equations that relate to previous years' teacher, school, family and community characteristics may lead to biased estimates; estimates that on average are incorrect, of the effect of class size on student educational achievement gains.

Other critics of the *Coleman Report* and much of the gain score literature that followed it were quick to point out, as we have already repeatedly stressed, that the average pupil/teacher ratio in a school is not the same as the class size that any individual student actually experiences. Many teachers, such as reading specialists and art, music, physical education and support teachers, do not have their own classrooms. This causes

the average pupil/teacher ratio in a school to understate the average class size. Within a school, class sizes also differ widely across students. For example, special education classes tend to be smaller than the classes in which most students are enrolled. Hence, the use of the average pupil/teacher ratio in a school rather than the class size in which a student is enrolled introduces measurement error into the variable of concern and makes it more difficult to "tease" out true class size effects from empirical models.

Finally, while perhaps the most ambitious social science data collection effort that had been undertaken up to the mid 1960s, the *Coleman Report* data did not always capture all of the dimensions of the teacher, school, family and community variables that in theory should be expected to influence student performance. For example, parents' education was inferred from their occupation codes, not collected directly. Similarly, per pupil expenditures on staff were collected but not total funds spent per pupil in a school district. Inadequate controls for the forces that in addition to class size that should be expected to influence educational outcomes make it more difficult to infer what the true class size effects are.

D. Nonexperimental Evidence on the Effects of Class Size on Test Scores and Drop-Rates in the United States

During the three-plus decades that followed the release of the *Coleman Report* literally hundreds of studies have been undertaken to analyze the impact of class size on educational outcomes. Many of the early studies used methodologies similar to the *Coleman Report*, in that they estimated equations like equation (2) in which the level of an educational outcome was the dependent variable. As time passed, increasingly the

studies took the form of estimating gain score equations in which the change in a test score measure between the current and last academic year (A(t) – A(t-1)) was specified to be a function of the levels of the school resources and teacher characteristics that a student was exposed to during the current academic year, the characteristics of the student's family during the year, the characteristics of the school community in which he or she was enrolled during the year, the characteristics of the community in which he or she lived during the year and a random error term. That is, they estimated equations of the form

(3) $(A(t) - A(t-1)) = b_o + b_1 r(t) + b_2 f(t) + b_3 s(t) + b_4 m(t) + e(t)$,

where b_0 is an intercept term and the other b's are vectors of parameters.

Some, but not all, of these studies included the level of the student's test score in the previous year as an additional explanatory variable on the right-hand side of the equation. The lagged value of the test score is included to capture the fact that the level of academic achievement that students start with in a year likely influences how much they learn during the year. However to the extent that the lagged test score variable is determined by the student's class sizes in earlier years, these prior years' class sizes are correlated with current class size, and these prior school years' class sizes are not included on the right-hand side of equation (3) (which they rarely are), the error term in equation (3) will be correlated with the current academic year class size and this will lead to a biased estimate of the effect of current academic year class size on the student's gain score.

Different studies use different types of data. Some use test score data on individual students within a single school district or state and actual class size data for

each student. Others use individual student data but use average class size data for students in that grade in each school. Still others use average scores for students in a grade level within a school and average class size for students in that school. Some do not have class size data and use the average pupil/teacher ratio within the school. Finally others aggregate even further and use average test score data by school district or state and the corresponding average pupil/teacher ratio. There are methodological problems associated with each of these types of data analyses, which we shall return to shortly.

Perhaps the most well-known participants in the "do school resources matter" debate are Eric Hanushek and Larry Hedges and his colleagues, who in a series of meta-evaluation studies, have presented conflicting views on the impact that school resources, including class-size, have on students' educational outcomes.[10] Hanushek has focused on the pattern of the signs and statistical significance of estimated class-size and pupil/teacher ratio variables in educational outcome equations that use either measures of the level of educational attainment, or the change in measures of educational attainment between two years, as the dependent variable. To be included in his summary, studies had to meet only minimal methodological specifications. To be specific, according to Hanushek (1999) "..the studies must be published in a book or journal (to assure a minimal quality standard); must include some measure of family background in addition to at least one measure of resources devoted to schools; and must provide information about the statistical reliability of the estimates of how resources affect student performance".

Table 1 provides a summary of his findings of the studies published through the mid 1990s. Panel A provides an overview of all of the studies, as well as of the studies

conducted using data from elementary and secondary students. In this table, if smaller class sizes are associated with more learning, this is reported as a positive effect, even though the underlying estimated coefficient of class size would have been negative. Hence positive effects mean that class size matters in the sense that smaller class sizes are associated with improved student performance.

The percentage of positive effects reported overall in these studies only marginally exceeded the percentage of negative effects (42% as compared to 38%). The percentage of estimated effects that were positive and statistically significantly different from zero using a two-tailed test at the 5% level (15%) was only marginally larger than the percentage that were negative and statistically significantly different from zero (13%). Restricting the comparisons to studies that used data from secondary schools yielded roughly the same conclusions. Moreover, a slightly higher percentage of estimated effects from studies that used data from elementary schools were negative rather than positive and about 20% of the elementary studies yielded statistically significant negative effects as compared to the 13% of these studies that yielded positive effects.

Panel B summarizes results from the same set of studies but this time the studies are grouped by the level of aggregation of the resource measure used. The studies classified under classroom use an estimate of the class-size in the class a student attended. Those classified under school use a measure of the pupil/teacher ratio in the school or the average class size in the grade in the school that the student attended. Those classified under district are studies that use school districts as the units of observation and are based on average pupil/teacher ratios in the district. Finally, county level studies use county level data as the units of observation and state level studies use state level data

and average pupil/teacher ratios in the state. Only for district and state level studies did the percentages of positive and of positive and statistically significant pupil/teacher estimated effects exceed the comparable percentages of estimated effects that were negative.

Most of the studies that are summarized in the table use the level of student achievement at a point in time as the outcome variable. We have already stressed why this type of model will not yield unbiased estimates of class size effects. So in panel C, Hanushek restricts his attention to estimates that came from models in which a gain score was the educational outcome variable. Only 78 estimates came from such studies and only a slightly higher percentage of these estimated pupil/teacher effects were positive and statistically significant than were negative and statistically significant.

Studies that use data that span more than one state are subject to the criticism that educational policies other than class size vary across states and that unless these policies are parameterized and included in the model (which they rarely are) the class size coefficients may be biased by their omission. For example, states that have smaller class sizes may also be the states that require students to take and pass competency exams each year. It may be the latter, rather than smaller class size, that leads to improved educational outcomes in those states.[11] So in the last row of the table, Hanushek restricts his attention to gain score equations that are based on data from within a single state. Here 52% of the estimated class size effects prove to be negative and the percentage of class size effects that are negative and statistically significant (13%) is more than 4 times the comparable percentage of estimated positive and statistically significant effects.

While Hanushek concludes from these studies that there is little reason to believe that smaller class sizes systematically improve student educational performance, Hedges and his colleagues view the same evidence somewhat differently.[12] They argue that *count tables* of the type that Hanushek has constructed are not sufficient to draw the conclusions that he has drawn for several reasons. The first is that count tables ignore the magnitudes of the coefficients in the studies being summarized. When they consider the magnitudes of the individual estimated coefficients, they find that both the median and average estimated coefficient often suggests that class size does matter.

The second is that estimated pupil/teacher ratio coefficients in individual studies may be statistically insignificant because of the small sample sizes used in the analyses or because it is difficult to obtain statistically significant estimates of effects that are relatively small in magnitude. However, there are formal statistical tests that allow one to compute combined significance tests of coefficients from a set of studies that seek to test the same conceptual hypothesis but that may differ in design or measurement.[13] When Hedges and his colleagues apply such tests to a sample of gain score studies that is similar to the sample analyzed by Hanushek, they find that they cannot reject either the hypotheses that smaller pupil/teacher ratios are associated with more learning by students or the hypothesis that smaller pupil/teacher ratios are associated with less student learning. However, when they "trim' their sample, by ignoring the 5% of estimated coefficients that were in each "tail" of the distribution of coefficient estimates, they find that they can reject the latter hypothesis.[14]

Who is correct, Hanushek or Hedges and his associates? In our view neither conclusion can be verified because neither meta analysis used stringent conditions for

deciding which non-experimental studies to include in their samples. Many, if not all, of the studies that they summarize have methodological problems, which almost guarantee that the estimates of the relationship between pupil/teacher ratio and student learning that each study reports will not be unbiased (on average correct). In some cases the direction of the biases can be signed (that is whether the estimated effect is an over or under estimate of the true effect). In other cases, they cannot. However, the fact that many of the individual coefficient estimates are likely biased leaves us wary of what the meta analyses really teach us.

To see why estimates of the effects of smaller class sizes on student learning obtained from nonexperimental data are often biased, consider the following simplified version of a gain score equation:

(4) $A(t)-A(t-1) = a_0 + a_1 x(t) + a_2 cs(t) + e(t)$

The change in test score that a student achieves between year t-1 and year t is specified here to be a function only of the size of the class in which he is enrolled in year t (cs(t)), a vector of other variables (x(t)), and a random error term (e(t)). The a_0, a_2 and the vector a_1 are parameters to be estimated. The estimated value of a_2 that one obtains is an estimate of the marginal effect on the gain score of changing class size by one student.

What are the circumstances under which estimation of equation (4) by the method of least squares analysis will provide us with an estimate of a_2 that is on average correct (unbiased). A number of assumptions must hold. The first is that we have good information on the class size in which a student was enrolled. If the class size variable is measured with error, the estimated effect of class size on gain scores will be biased. The direction of the bias will depend upon how the measurement error is correlated with the

error term in equation (4). For example, if the measurement error is random, it is well known that the estimated coefficient will be biased towards zero.

There is considerable measurement error in the class size variable in most studies. Most use measures of school, school district, county, or statewide average pupil/teacher ratios rather than the actual class size in which a student was enrolled. This leads to two types of measurement error. On the one hand, an individual's class size does not necessarily equal the average class size in the broader unit. Part of the variation may be due to random factors such as classroom size limitations. However, part may be due to the type of class in which a student is enrolled. For example, classes for students classified as needing special education are often smaller than regular classes.

On the other hand, as we have already noted, the average pupil/teacher ratio in a school understates the average class size in the school. The reason for the latter is that the many teachers do not have their own classes. For example, support teachers, reading teachers, and specialists in elementary schools (art, music and physical education teachers) do not have their own classes. Whether these two types of measurement error should be considered random is not obvious.

Another key assumption is that the error term is uncorrelated with either class size or any of the other variables that are included on the right hand side of equation (4). One reason that the error term might be correlated with class size is that the size of classes in which students are enrolled may be systematically correlated with how much they are expected to learn during the year. For example, if a school groups its students in a grade by their "ability level" and believes that lower ability students will learn less during the year, to compensate for this it may assign lower ability students to smaller class sizes.

Thus a smaller expected gain score will cause a student's class size to be smaller than would otherwise be the case. This builds in a spurious positive correlation between class sizes and gain scores (the brightest students who are expected to learn the most are in the largest classes) and will cause us to underestimate the true effect of reducing class size on student learning.[15]

To control for this problem some studies try to include among the other variables in the model variables that are expected to influence class size. For example, if expected gains score growth depends upon an individuals test score level in the previous year, these studies would include the student's lagged test score among the other variables included in the estimation equation. While their inclusion controls for the "endogeneity" of class size, it introduces another problem which we have already discussed, namely that the lagged test score depends upon variables such as previous year's class size, which typically are not included in the model. To the extent that lagged and current class size are positively correlated and lagged class size affects lagged test scores, excluding lagged class size from the right hand side of equation (4) will lead to biased estimates of the effect of current class size on gain scores. This is but one example of the "omitted variable problem"; we return to this statistical problem in a moment.

A few researchers have tried to control for the endogeneity of class size measures by exploiting variations in class size that occur for reasons independent of student's expected gain scores. For example, if one is estimating the gain-score class-size relationship using data from small schools that have only one class in each grade in a school, variations across schools in class sizes in a grade will reflect only variations in population sizes in the grade across the school areas[16]. While this strategy may help to

yield unbiased estimates of class size effects for small schools, it does not enable one to more generally estimate the impact of class size on gain scores. Other authors attempt to overcome the endogeneity problem by using *instruments* for class size[17]. This requires them to find a variable that is highly correlated with class size but that itself is not correlated with the error term in the model.

More generally, estimates of class size effects may be biased because the underlying model excludes from the vector of other variables x(t) in equation (4) variables that have an independent effect on educational attainment and that are correlated with the class size that a student experiences in a year. Our discussion of what an ideal estimating equation would include emphasized that educational outcomes may depend upon a whole vector of teacher characteristics including education, academic ability, experience and a whole vector of school level characteristics, such as the number and quality of support staff and specialists, the quality of the principal, and the level of resources devoted to educating each student. Most educational production function studies include some subset of these variables but rarely are all of them included.

For example, several studies have concluded that teacher verbal ability is an important determinant of how much students learn.[18] Most nonexperimental data sets that have been available to researchers who are trying to estimate class size effects do not contain measures of teacher verbal ability. To the extent that schools or school districts that have small class sizes also hire teachers with high verbal ability, the omission of teachers' verbal ability from estimating equations will cause the effects of class size on educational outcomes to be overstated in absolute value. Similarly, to the extent that teacher subject matter competencies influence learning, if school districts with smaller

class sizes also hire teachers with higher levels of subject matter competencies, the omission of teachers' subject matter competencies from estimating equations will cause the effects of class size on educational outcomes to be overstated in absolute value.[19]

A second type of omitted variable bias arises from the endogeneity of where students attend schools. Students are not randomly distributed across public schools in the United States. Rather where parents choose to live determines, to a large extent, the nature of the schools that their children choose to attend. Suppose that parents who value education highly and who have the resources (income and assets) to afford to live in areas where schools are thought to be "good" and that have small class sizes locate in these areas. Suppose also that these parents "invest" heavily in their children's education outside of the school. This may take the form of expensive preschools, of having computers in the home, of working on homework with children, of emphasizing to children the importance of learning and of staying in close contact with teachers so the parents will know what problems their children are having in school. If one observes that children who attend "good" schools with small class sizes learn more than students who attend schools with larger class sizes, part or all of the differential investment may be due to the parents financial and time investments in their children's education.

Unless the vector of variables x(t) in equation (4) fully controls for home investments of the type described above, estimates of the effects of class size on student learning will be biased. Most educational production function studies attempt to control for these factors but the data that researchers use do not always permit them to capture all dimensions of home investments. For example, some data sets to not include accurate

measures of parental income, assets or education and rarely do the data sets actually contain information on the home investments that parents made in their children.

More generally, it should be clear to the reader from our discussion above that the accurate estimation of class size effects is part of a larger set of questions concerning the estimation of classroom and school effects (Raudenbush, & Willms, 1995; Willms, & Raudenbush, 1997). These questions include, for example, whether certain teaching strategies or approaches to classroom discipline are more effective than others, whether ability grouping or "streaming" has positive or adverse effects, or whether parental involvement in school governance improves academic achievement. Indeed many of the most important and interesting policy questions concern the effects of factors at various levels of the schooling system on children's outcomes. Researchers have struggled with how to obtain accurate estimates of classroom and school effects for many years. There are at least five basic problems- *selection bias, confounding variables, low variation in the independent variables, cross-level interactions and latency*- that make estimation of these effects difficult. We briefly discuss each problem in turn.

The most critical problem is selection bias. Children from higher socioeconomic backgrounds on average begin school better prepared to learn and receive greater support from their parents during their schooling years. In many instances, the policy variable of interest is correlated with the students' family backgrounds. Therefore, the accuracy of the estimated effects of the policy variable will depend crucially on how well the researcher is able to control for family background factors. For example, in a recent study of over one thousand primary schools in Latin America the correlation between the pupil/teacher ratio in the school and the socioeconomic level of students in the school

was about -.15. That is, schools that enrolled advantaged students on average had lower pupil-teacher ratios. Moreover, the schools enrolling students from more affluent backgrounds tended to have better infrastructures, more instructional materials and better libraries. The correlations of these latter variables with school-level socioeconomic status ranged from .26 to .36 (Willms, & Somers, 2000).

Similarly, at the classroom and school levels there are usually other factors, or confounding variables, that are correlated with the policy variable of interest. For example, in many school systems the most qualified and experienced teachers may be attracted to and remain in schools with smaller classes. So the problem of selection bias may be exacerbated by these confounding variables.

A third problem is that many relevant policy variables may not vary much across schools settings. The greater the variation in a variable across observations, the more accurately one can estimate the effects of that variable on an educational outcome. Class size is a good example, because in most states there is a relatively small range of class sizes among schools at a given grade level, which makes it difficult to disentangle the effects of class size from other factors.

The effects of certain policies or practices may vary for different types of students. For example, in small classes teachers may be able to spend more time with students who are struggling with the content of a lesson, or cope with the disciplinary problems presented by children with behavior disorders. Thus there may be a greater benefit of small class sizes for students with low cognitive ability than for more able students and for classes that contain some children with behavior disorders than for classes that have no such children. When variables at two different levels, such as

classroom (teacher) and student interact, this interaction is referred to in the literature as *cross-level* interactions. *Same-level* interactions are also possible; for example, small classes may yield greater gains for students when experienced teachers are present than when inexperienced teachers are present, or visa-versa.

The final problem concerns the potential *latency* of intervention effects. For example, recent neurobiological research has shown that brain development from conception to age one is rapid and extensive and that there are critical periods, especially during the first three years of life, when particular areas of the brain are "sculpted" (Cynader, & Frost, 1999). At the same time, longitudinal studies have shown that intensive interventions aimed at increasing stimulation and providing parental training for low socioeconomic status families can improve life trajectories of economically deprived children. There is a latent effect that is not immediately evident during the period of intervention.

One could similarly hypothesize that children educated in small classes during the elementary grades are more likely to develop working habits and learning strategies that would enable them to better take advantage of learning opportunities in later grades. Such latent effects are difficult to estimate because the initial effects of the intervention, in this example working habits and learning strategies, may be associated with constructs that are difficult to precisely define and measure. We may also not be able to predict in advance when these constructs should be expected to begin to influence measured outcomes, such as the growth in test scores.

A related, but different problem, is that an intervention may not fully "take hold" until two or three years after it is introduced. For example, as we shall discuss in more

detail in section IV, it may well be the case that the effects of smaller class sizes crucially depend upon teachers altering the way that they teach. However, after a reduction in class size, it may take two or there years for teachers to begin to modify their teaching practices, if they do so at all, so that the potential benefits from the smaller class sizes can be realized.

Section III, below, will discuss "true" experiments, in which students are randomly assigned to treatments. In the case of class size experiments, students are randomly assigned to classes of varying sizes. This approach has several advantages, because it explicitly removes the threats of selection bias or confounding variables. Also, if the treatment contains a wide range of class sizes, the third problem discussed above is less of an issue.

However, true experiments have their own problems. People who are being studied often react differently than if they were in natural settings. For example, in a class size experiment, the parents of children who are assigned to large classes may try to compensate for their children's placement by providing their children with after school tutoring. Such "compensatory rivalry" may distort the experimental results. Parents of children assigned to large classes may also react with apathy or hostility, in some cases even trying to subvert the experiment. Educational psychologists call this phenomenon "resentful demoralization" (Cook, & Campbell, 1979).

Some parents may try to have their children reassigned to the smaller classes and there could be jostling among teachers for more favorable classroom assignments. Student attrition rates may differ between students in the treatment and control groups, resulting in a problem that is similar to selection bias. Perhaps the biggest shortcomings

of true experiments though relate to their costs. They are often too expensive to implement on a large scale making it difficult to generalize across grade levels and settings. Moreover, when the number of classes and schools being studied is small, it becomes difficult to reliably assess cross-level interactions or same-level interactions at the classroom and school levels.

In some respects, therefore, there is no substitute for an "on the ground" description of the relationship between children's schooling outcomes, family background and school policies as they occur in natural settings. However, to achieve reasonable estimates of class size, the researcher must ameliorate the threats to validity described above with strong research designs and statistical methods.

The most common approach to statistical control is multiple regression analysis, as typified by equations (2) and (3) above. An important assumption underlying traditional multiple regression analysis is that the observations in the sample are independent of each other. For example, if the data are data for individuals, any one individual's data are assumed not to be systematically related to the data for any other individuals. However when students from the same classroom are in the sample, this assumption is violated.

Researchers have long debated whether the appropriate level of analyses to conduct class size effects studies was the pupil, classroom or school level. If individual data were to be used, they worried about the assumption of independence when multiple students from the same classroom were included in the sample. Fortunately, recent advances in statistical theory and computing have permitted statisticians to develop *multilevel* regression models that can include data at different levels without violating the

assumption of statistical independence (Goldstein, 1995; Raudenbush, & Bryk, 1986). These models allow an analyst to explicitly model the effects of variables describing school and classroom level policies and practices on students' educational outcomes. Recent studies using a multilevel approach have yielded findings that are remarkably consistent with Project STAR, a large-scale true experiment that we discuss in a later section. Readers interested in learning more about these models and their application to class size issues will find an introductory discussion in the appendix.

E. Nonexperimental Estimates of the Effect of Class Size on the Subsequent Labor Market Earnings of Students in the United States

Economists have long argued that a major role of elementary and secondary education is to prepare students to participate in the labor market. Thus, they believe that the effectiveness of resources devoted to schools should also be measured by the impact of school resources on the earnings of their graduates, as well as by the impact of the school resource on students' test scores. In an important study, David Card and Alan Kruger (henceforth CK) used individual-level earnings data for white men born between 1920 and 1949 from the 1980 *Census of Population* and matched these data up with various school quality measures that existed in each individual's states of birth during the years that he was enrolled in school to see if characteristics of the school systems in their birth states influenced the rate of return that the men earned on their educations.[20]

CK found strong evidence that measures including the average pupil/teacher ratio in the state, the average education level of teachers in the state, the average teacher salary relative to mean earnings in the state (a measure of the attractiveness of teaching as a

profession in the state) and the average length of the school year in the state, were all related to an individual's current earnings, holding constant his age and their education levels. Thus, for men between the ages of 31 and 60 in 1980, CK found evidence that smaller average pupil/teacher ratios during the years they were educated were associated with higher subsequent earnings. Pupil/teacher ratios did matter.

CK were forced to use statewide average characteristics of the educational system that existed while the men in their study were being educated because the individual data from the 1980 *Census* do not contain any information on the characteristics of the schools that individuals who were surveyed in the Census attended. In contrast, in a series of papers, Julian Betts used data from the *National Longitudinal Survey of Youth* (*NLSY*), a nationally representative sample of individuals who were ages 14 to 24 in 1979 when they were first surveyed (Betts, 1995; Betts, 1996). The *NLSY* contains information on characteristics of the high schools that the students attended, including average class size. The sample was resurveyed periodically and by 1990, they were ages 25 to 35. Given that 1990 earnings data were available in the *NLSY*, Betts was able to come close to replicating CK's analysis, using the young men's actual schooling variables rather than the statewide averages that CK used.

When Betts did this, he found no evidence that any of the school measures contained in the *NLSY* data, including class size had any statistically significant effect on the young men's earnings. However, when he replaced the actual value of the school variables in the school that each individual attended by the average value of the same variables for all schools in the state, he found that many of the statewide average variables were statistically significantly different from zero at conventional levels of

significance, which suggested that school resources, including class size did affect earnings. Put another way, Betts found that statewide average characteristics of schools appear to be associated with individuals' subsequent earnings but that the characteristics of the schools the individuals themselves attended were not. He concluded, as we have done earlier, that the estimated impact of the statewide average class size variable might reflect other differences that exist across states rather than class size differences per se.

Needless to say, the debate rages on about what the difference in findings between the two studies mean. CK have argued that small sample sizes in the *NLSY data* (in the range of a few thousand rather than the hundreds of thousands of observations available in the 1980 *Census* data) make it difficult to observe statistically significant effects (Card, & Kruger, 1996). They also argue that the use of state averages for school characteristics is preferred to the actual data on individual school characteristics because the latter are measured with considerable measurement error. This measurement error arises because the reported school characteristics data only reflect the characteristics of the school that a student attended at a point in time. School characteristics, such as class size vary over time for a student at a given school and many of the individuals in the sample attended more than one school in a given district and more than one school district during their elementary and secondary school careers. Finally, CK suggested that the relatively young age of the individuals in the *NLSY* data make it difficult to directly compare their results and Betts's. However, other more recent work suggests that the CK class size findings are very sensitive to the functional forms they assumed in their research and are not robust to functional form assumptions (Heckman, Layne-Farrar, & Petra, 1996).

Our overall conclusion is that the literature on the effect of class size on subsequent labor market earnings suffers from many of the same problems as the literature on class size effects on test scores and schooling levels. There are simply too many statistical problems relating to measurement error and omitted variables problems in these nonexperimental studies to place great faith in any of the findings. Studies that find statistically significant effects of class size on labor market earnings tend to use aggregate statewide averages for these variables. We are not sanguine that using such an aggregated approach allows the researchers to conclude anything about the effects of class size, per se.

III. Experimental and Quasi-Experimental Studies

A. The Advantage of Experiments

The most fundamental question about class size is does it affect student learning? This most basic of questions has received plenty of attention in research. In principle, the best way to determine this basic relationship would be an experimental design in which students are randomly assigned to classes of various sizes and then followed over time on a number of outcome measures. Randomization means that otherwise identical students receive a well-defined treatment, so group differences can later be reasonably attributed to that treatment. This kind of design is generally regarded as the 'gold standard' and the best way of determining program effects. However, in order for correct inferences to be made, the experiment has to be conducted in a careful and controlled manner and on a large scale – the treatment needs to be clearly defined and measured; the randomization accurately conducted (i.e. no bias in assignment to control and treatment groups); little

crossover (i.e. assignments have to stick); attrition be kept to a minimum, so that randomization and large enough samples are maintained; individuals followed for a long period of time and data collected on multiple outcome measures. The experimental design and outcomes also need to be independently verifiable. In addition, experiments may have limited external validity and therefore be of limited use for policymakers considering expanding the treatment to a broader scale (i.e. beyond the conditions that prevailed during the experiment). Therefore although well-done experiments can shed light on the basic effects of an intervention, they may not always be useful for policymaking purposes.

For a number of reasons beyond the scope of this article –ethical concerns, inadequate research funding, etc. – experiments are very rare in education. Large-scale long-term experiments with independent evaluations are even rarer. To the extent that experiments are conducted at all they tend to be very small scale (e.g. in one school or with a few students), and rarely evaluated in a manner that would give one much confidence in the results. Class size is the exception to this rule, as the results of a demonstration conducted in Tennessee in the 1980s has attracted widespread attention. Project STAR (Pupil/Teacher Achievement Ratio) is believed by many in the education research community to provide the most definitive evidence to date on the class size issue.

STAR has been called by Mosteller (1995) “one of the greatest education experiments in education in United States history”. Some have argued that the results of this experiment call into question the validity of the entire body of non-experimental work (Grissmer, 1999) and that “they eclipse all of the research that preceded it” (Finn, &

Achilles, 1999). However, we think this position *overstates* the importance of the results. As Hanushek (1999) has correctly pointed out:

> "even medical experiments with well designed protocols and well defined treatment programs frequently require more than one set of clinical trails to ensure valid and reliable results. Social experiments, which tend to be more complex, are very difficult to design and implement, making it even less likely that a single trial will provide definitive answers".

Here we briefly review the design of STAR, its main findings, and the points of disagreement among analysts. It is particularly important to be aware of the specific design features and conditions under which STAR was implemented – both because these affect whether the results are believable and also because they impact the validity of inferences drawn for actual class size reduction policies that have been implemented subsequently.

B. Tennessee s Project STAR (Pupil/Teacher Achievement Ratio) Experiment *Design*[21]

STAR was a state sponsored $12 million "demonstration" that began in 1985. Students entering kindergarten were randomly assigned to one of three treatments – a class of 13-17, a class of 22-26, or a class of 22-26 with a full time aide- for four years after which they returned to a regular size classroom. A new teacher was randomly assigned each year. Treatments were maintained throughout the school year and there were no other interventions from the state (e.g., no special teacher training or new curricula materials). Students and teachers were randomly assigned within a school,

"thus controlling for differences between schools that might be important in explaining student achievement, such as differences in the populations served, differences in per pupil expenditures and instructional resources, and differences in the composition of school staff" (Finn, & Achilles, 1999). The state paid for the additional teachers and aides and provided that no student would receive any less service than would normally be provided by the state (Nye, 1999).

All districts in the state were invited to participate but participating schools had to agree to the random assignment of teachers and students among the three class types as well as data collection for the evaluation (including standardized achievements tests for students at the end of each grade). Hence a minimum of 57 students (13 in a small class, 22 in a regular class and 22 in a regular class with an aide) was required so that very small schools were systematically excluded. 180 schools expressed interest (in 50 of the state's 141 districts but only 100 met enrollment requirements. Selection among these schools was based on a requirement that at least one school be included from each district that had volunteered, and that there were schools in rural, urban, inner city and suburban districts and three regions of the state in the initial year. The median class enrollment before the demonstration was 24; after it was 15. The final sample included 128 small classes (1900 students), 101 regular classes (2300 students) and 99 regular classes with an aide (2200 students), about 6000 students in 329 classrooms in 70 schools and 46 districts in the first year, and almost 12000 students over the 4 years intervention due to the addition of new students. As Hanushek (1999) has pointed out the samples used in published research are typically smaller due to missing information.

C. Results from the STAR Experiment

There have been numerous analyses of the Tennessee STAR data and not all authors agree on the results.[22] A team based in the state originally conducted an evaluation, but in recent years numerous other academics have investigated the data as subsequent longitudinal outcome data for students in the original demonstration have been collected. Finn and Achilles (1999, p. 98) summarize their view of the findings as yielding "an array of benefits of small classes, including improved teaching conditions, improved student performance and after the experimental years improved student learning behaviors, fewer classroom disruptions and discipline problems, and fewer student retention".

Statistically different achievement differences were found between students in small classes and the two other groups (and no differences between classes with aides and without). For all students the difference was around a fifth of a standard deviation in student achievement, the gap generally appearing by first grade. There are significantly larger (by two to three times) effects for minority students, a finding replicated by Krueger (1999) (although the overall effects are smaller).[23] One way to think about whether from a policy perspective the effect for minority students is a large or small difference is to consider that the typical gap observed in achievement between minority and non-minority students on most standardized tests is about one full standard deviation.

The differences between the small and large class groups appeared to persist in to the upper grades. In grades 5-7, after the students had returned to regular size classes, achievement effects in a wide range of subjects persist in the one-tenth to one-fifth of a standard deviation unit range. Subsequent analyses (e.g., Krueger, 1999; Krueger, &

Whitmore, 2000; Nye, 1999; Finn, Gerber, Achilles, & Boyd-Zaharias, in press) have generally found that effects are similar and long lasting, even after utilizing statistical models that control for various covariates such as school effects and teacher characteristics (Krueger, & Whitmore, 2001). Differences between minority and non-minority students have been examined only through grade 4; these show that the early benefits of class size for reducing minority disadvantage persist, but do not expand, after the class size experiment has ended (Finn, Fulton, Zaharias, & Nye, 1989). The most recent analyses indicate that the more years students spend in small classes during grades K-3, the longer the benefits for achievement last during grades 4-8 (Finn, Gerber, Achilles, & Boyd-Zaharias, in press). At no point were differences found that support the aide treatment.

Table 2 contains some specific estimates of the small class advantage by grade level in mathematics and reading, found in the STAR study. There is some disagreement over the precise size of the small class effect since "they are dependent on student characteristics, the length of time and which grades were spent in small classes, the test and units of measure used to measure the effects and whether the focus is on short-term or long-term effects" (Grissmer, 1999). Most notably there is some division over duration and timing effects, something that is critical for policy – how big the effects are for students who are in reduced classes for several years and when the effect occurs. For example, Krueger (1999) estimated that students in four years of smaller classes had about a .19-.25 standard deviation unit advantage, while Nye, et al., (1999) suggest it was almost twice this (.35-.39), with the former suggesting most of the boost comes in the

first two years of exposure and the latter effects only from sustained exposure. Clearly it is important to know which of these researchers is correct.

This discussion raises the conceptual point of what one might *expect* to see from a class size treatment. In other words, is there a link between exposure to smaller classes over a time period and student achievement – would two years in a smaller class rather than one, for example, double the first year advantage in achievement, or simply maintain it? If the latter is the case, is that second year needed or would the advantage that appeared after the first year persist regardless of whether the student were in a large or small class in the second year? The general pattern of STAR results reported in standard deviation units in the table suggests that effects are not cumulative but they do persist. Hanushek (1999) argues that this "ignores the fact that one would expect the differences in performance to become wider through the grades because they continue to get more resources (smaller classes) and that should keep adding an advantage". He plausibly suggests that the results are consistent with a one-time effect and shows that the effect for those in the experiment all four years is smaller compared to annual samples whereas a cumulative story would suggest a larger effect should be present. By contrast, Finn and Achilles (1999) argue that stable effect sizes are partially a spurious result due to test publishers scaling procedures. They recast the STAR results in terms of "grade equivalents" and show effects *increasing* with each grade, although this is a controversial procedure.

There are various issues surrounding the results of the STAR project, some of these are related to the design and implementation, some are related to various methods of evaluating the data. The most vocal critic has been Hanushek (1999), although others

have also raised concerns (Hoxby, 2000). The most strident defenders of the validity of the study have been members of the original study team (Finn, Achilles, etc.) and the authors of some newer studies (Nye et al., Krueger).

D. External Validity and the Hawthorne Effect

An experiment and its findings may be very dependent on the conditions under which it was conducted. In the case of class size in Tennessee for example, there was an ample supply of qualified teachers and no facilities shortages, and additional resource costs were borne by the state, such that implementation could proceed smoothly, without the need for reallocation of other resources. This may not be the case for a large-scale class size reduction policy (see below on California). In addition, the results apply strictly only to the particular population of students participating – and the characteristics of the students in the experiment were different than those of average Tennessee and national students (poorer, more minority) (Grissmer, 1999).

One of the design features of STAR was the non-random selection of *schools*. Although this is unlikely to bias the estimated program effects given within school randomization of teachers and students – and in fact studies that have estimated statistical models that include school effects have confirmed this - this is clearly a problem for the generalizability of the studies findings to other settings. Of course even within a school some small classes produced bigger effects that others – so other factors such as teacher quality, and peer composition, are also likely to be important factors. Given the difficulty of measuring these concepts the precise interaction of the various factors is not clear. Again, this is an important omission since it may be that certain kinds of teachers, or

enhanced teacher professional development, or particular groupings of students, can add or detract significantly from class size effects.

The actors in the experiment are aware of it. It is a commonly observed that even if a policy has no effect, people may behave differently if they are being evaluated. In addition, Hoxby (2000) has pointed out that since the success of experiments often determines whether a policy is implemented universally, it may be that "the actors have incentives to make the policy successful that they would not ordinarily have." Although there is no explicit evidence that these effects occurred in the case of STAR it is quite possible that they occurred and Hanushek (1999) has suggested that the significant reassignment of students across treatment groups that occurred in STAR and predominately in the direction of small classes "clearly indicate that school personnel reacted to participant desires in this nonblind experiment".

E. Randomization, Attrition, Crossover

For various reasons, the process of randomly assigning students and teachers in an experiment like STAR is not necessarily straightforward in practice. Student mobility due to parental pressure (e.g., a vocal parent requesting assignment to a particular treatment) and administrative strategy (e.g., principals assign teachers to classes where some kids are thought most to benefit) can affect in often subtle and unobservable ways whether there is true randomization at the outset of the experiment – and it is from this that all subsequent conclusions about program effects follow. Knowing whether in fact the randomization was well executed initially is difficult to determine without good data, and this is difficult to verify in the case of STAR (Hanushek, 1999). For example,

although Krueger (1999) has shown that there is no evidence of non-random teacher assignment, this is based on available measures of teacher characteristics (education and experience). Although Grissmer (1999) suggests that it is unlikely that unobserved teacher quality would therefore be non randomly distributed, the correlation between these measures and teacher quality or effectiveness in the classroom may be very weak, and it seems quite plausible that these factors could have been used in the teacher assignment process. It is also known that unobservable teacher traits have a significant role in explaining student achievement while observable characteristics do not (Goldhaber, & Brewer, 1997; Hanushek, Kain, & Rivkin, 1998). For students, no pretest of entering students was given (admittedly problematic in kindergarten, though not for older students added later) that would permit a true check on random assignment.

Although the goal was for students to remain in the same treatment for four grades, some changes actually occurred due to attrition and late entering students. An additional 6000 students entered the sample by the end of the demonstration. And although students entering a school in later grades were randomly assigned, they had likely come from schools in which they had experienced larger classes, and late entrants generally had lower test scores. In addition, as Hanushek (1999) has noted, some of those entering in first grade may not have had kindergarten at all since it was not compulsory in Tennessee at the time. Grissmer (1999) argues therefore that "the reported experimental effects should not be compared across grades since the small class sample past kindergarten changed in each year, containing children who had spent different numbers of years in small classes".

In addition, there was considerable attrition in the STAR groups. Of the initial experimental group, 48% remained in the experiment for the entire four years (Hanushek, 1999). The annual attrition rates are 20-30% and are not random in the sense that stayers look different than movers (Goldstein, & Blatchford, 1998), although reanalyzes of the program effects generally suggests that attrition patterns were similar across small and large classes (Nye, 1999; Krueger, 1999; Goldstein, & Blatchford, 1998). In fact, Nye et al (1999) found that "the students that dropped out of the small classes actually evidenced higher achievement than those who dropped out of the larger classes, suggesting that the observed differences in achievement between students who had been in small and larger classes were not due to attrition". Thus although one cannot be certain that biases did not creep into the STAR demonstration as it proceeded, they do not appear to threaten the basic conclusions of the study. Of more concern, perhaps, is the considerable switching that took place between control and treatment groups in the sample. Hanushek (1999) shows that the flows were larger in the direction of smaller classes and Word et al. (1990) document that the crossover was not random. Nye et al. (1999) attempt to test this problem by analyzing the small class effect using the initial assignment of students but these results prove identical to when actual assignment is used, so it is not clear how far there is a problem with the findings. Finally, Hanushek (1999) has raised the issue of non-participation in test taking, which applied to as much as 10% of the sample. The percentage of students excluded appears to be the same in both treatment and control groups, but again it is possible that there was non-random non-test taking across the groups.

Taken overall, although the results of one experiment must be treated cautiously, the STAR study results do appear to be reasonably robust in the sense that there is a statistically significant effect of being in a class of 14-17 rather than a class of 23 in an environment of ample teachers and facilities, and this advantage appears to persist well into upper grades after students have returned to larger classes. Although the advantage is persistent, it is not cumulative. That is, the advantage that emerges in kindergarten and first grade does not become larger, even when small classes are maintained in second and third grades. However, the early benefits of small class size, once established, persist at least through the upper elementary grades. Although there are a few legitimate concerns about the design and implementation of the study that could have been rectified with better data, it is very difficult to document that these cause considerable threat to the basic findings. We view the bigger threat as one of correct inference and interpretation of the results for policy, which is an issue of external validity rather than the technical merits of how internally valid the experiment was in practice.

F. Other Experiments and Quasi Experiments

There have been over 100 other small-scale experiments and quasi experiments that have focused on class size. These have been extensively reviewed elsewhere.[24] Such syntheses generally conclude that there is some evidence of a positive relationship between class size reductions and student achievement (Nye, & et al., 1999; Finn, & Achilles, 1999), particularly in the early grades for classes below 20, and for at-risk students, though the precise magnitude and linearity of effects[25] is open to debate. Like much of the non-experimental literature these findings tend to be treated somewhat

cautiously. They are typically poorly designed (true randomization is rare), of short duration, of small scale, and not subject to rigorous or independent evaluation.

In recent years, class size reduction has been adopted and implemented as a proactive policy in various states and districts and several good evaluations have been conducted of the outcomes of these policies. They do not, however, have the key randomized student assignment feature of STAR but nonetheless provide important supplementary evidence on what might be the effects of reducing class size. We discuss the results of two such evaluations – Wisconsin's SAGE (Student Achievement Guarantee in Education) program and California's CSRP (Class Size Reduction Program).

G. Wisconsin s SAGE Program

In 1996, Wisconsin began a five-year pilot study called SAGE – Student Achievement Guarantee in Education. Details can be found in Molnar, et al. (1999). This program includes several interventions adopted together, one of which is reducing the class level pupil teacher ratio to 15 students – such that some schools chose to have 30 students and 2 teachers -in K-3 (two others are extended morning and evening school hours and "rigorous" curricula and development of a system of staff development, but these have not been widely implemented). The policy was phased in, and is highly targeted towards schools that have 30% or more students below the poverty level within districts that have at least one school with at least 50% of their students below the poverty level. The scheme was funded initially at a flat amount of $2000 per low-income student in a SAGE class.

The evaluation involves a comparison between first grade students in SAGE schools and a group of comparison schools that have similar family income, achievement, enrollment and racial composition. The published results from the first two years suggest that the first grade SAGE students gained about 0.2 standard deviation units across tests as a result of being in classes of between 12 and 15 (average of 13.47 in 1997-8) compared to classes between 21-25 (average of 22.42 in 199708) in comparison schools. Interestingly the gains for African American students were substantially greater than for white students. Based on attitudinal survey data the evaluation team suggested "teachers in SAGE classrooms have greater knowledge of their students, spend less time managing their class, have more time for instruction, and individualize instruction using a primarily teacher-centered approach (Molnar, & et al., 1999). The SAGE experience is interesting because it has imbedded a relatively well designed evaluation; it is also an example of a highly targeted program, in contrast to the wholesale 'one size fits all' approach take by California (see below). However, it is not clear why the results for SAGE deserve particular attention relative to other small scale programs (SAGE involved just 14 schools in the comparison group in 1997-8, for example), save that it is more recent, and that it tends to confirm the STAR results.

H. California s CSRP (Class Size Reduction Program)

In 1996, California embarked on a massive statewide implementation of class size reduction in grades K-3 at a cost of well over $1.5 billion annually and involves 1.8 million students, beginning with 1st and 2nd grade, reducing class size from an average of 28.8 (maximum of 33) to a maximum of 20. This was a very different treatment from

what had occurred in Tennessee in STAR or has been implemented in Wisconsin through SAGE, particularly in that this was done in the context of limited facilities due to growing enrollment and an existing shortage of qualified teachers. Unfortunately, from the standpoint of a "clean" evaluation, the CSRP case is very problematic – not only was there no randomization, but the policy was implemented across the state very rapidly, and there was no state testing system in place initially. There was also no evaluation design in place before the program was implemented. While CSRP therefore provides some useful information on some aspects of the translation of the idea of smaller classes into large-scale public policy, it is less useful in pinning down the student achievement effects of class size.

The program has been evaluated most comprehensively by a consortium of leading research institutions (Bornstadt, & Stecher, 1999; Stecher, & Bornstadt, 2000) as well as by a handful of other researchers. As these authors point out, the analyses of the class size effects on achievement are limited by the late start of the evaluation (the state decided to fund one after the program was in place) and given the very rapid adoption of the policy there was only limited variation in class sizes in place in third grade (the only grade with reduced classes that has statewide test score data) and no pre-test or control group. This poses a real challenge to researchers at measuring class size effects. The initial approach of the CSRP consortium has been to analyze differences in third grade test scores for students in classes 20 or smaller and those greater than 20, using fourth grade score as a proximate control for pre-existing differences between students at a school. (Various other observable characteristics of students and schools – e.g. demographic information - are also utilized.) While this is a reasonable approach it is

certainly not ideal. The basic finding in both first and second year evaluations is that there is a small, statistically significant achievement advantage in reading, writing, and mathematics (but not in spelling) for all students in small and large classes in the third grade of about 0.05-0.10 standard deviation units. Interestingly the effect size does not appear to vary across different types of students based on race/ethnicity or poverty status.

There are several potential explanations for why these results are "disappointing" in the sense that results are smaller than those seen in STAR. Obviously the intervention itself was different – class sizes were reduced from a little less than 30 to 20 rather than to 15. Most notably the implementation of a large scale policy is very different from a small controlled experiment – many districts in California were already facing teacher and facilities shortages, and implementing CSRP required a considerable reallocation of resources, including the hiring of many uncredentialed and inexperienced teachers and the use of portables or non-classroom space for the newly created classes (Bornstadt, & Stecher, 1999). Given data limitations it is not possible to discern how these consequences affected student achievement but they surely must have had some impact. Moreover, poor school districts and districts with a high percentage of minorities had greater needs for new teachers and facilities and found it difficult to obtain them. Urban, rapidly growing, poor, high minority districts have very little space to build new classrooms, and have a hard time attracting qualified teachers. Consequently CSRP has widened the gap among schools in their resources, including the proportion of teachers who are fully credentialed (Reichardt, & Brewer, 1999). Poorer districts also had larger class sizes before CSRP and consequently need to do more to get class sizes down to 20

(and they only received state funds if class sizes were 20 or below).[26] The program is also in its early stages and it may take some time for effects to be seen.

Over the next few years, it is likely that some new research will appear on class size based on experimental and quasi-experimental designs. The STAR data are continuing to be analyzed by scholars not associated with the original research team, as the original STAR students continue to be followed through high school. Both California and Wisconsin programs, and their evaluations, are still in relatively early stages, and since many other states have adopted class size reduction policies of one sort or another (Parrish, & Brewer, 1998; Brewer, & et al, 1999), there is likely to be additional evidence from these attempted implementations of smaller class sizes.

IV. Why Does Class Size Matter? Inferences from Existing Research

There are many reasons why smaller classes might contribute to higher achievement, including better teacher contact with parents and more personal relationships between teachers and students. However, because classroom instruction is the most powerful aspect of schooling for achievement, the effects of class size on achievement are most likely to occur if class size is linked to instruction (Barr, & Dreeben, 1983). This linkage could be manifested in two ways. First, class size reductions may change what teachers do. That is, teachers may teach differently in smaller classes. If the changes were beneficial for students (e.g., more frequent assessments, more writing, more discussion, more help for individual students, and so on), then achievement would rise. The direct cause of this achievement increase would be instructional improvements, and class size would be the indirect cause. Second, even

if teachers do not change instructional practices, certain practices may work better in smaller classes. For example, students may pay better attention when there are fewer students in the room. Similarly, teachers who use a lot of small group work may find their instruction is more effective in smaller classes, because fewer students remain unsupervised while the small group meets with the teacher. In these instances, teachers could carry on the same practices, but achievement would rise in smaller classes because the same instruction would be more effective. According to this account, class size and instructional practices would interact to affect student achievement.

All teachers face a fundamental problem of establishing and maintaining order in their classrooms. Most teachers, at least in the United States, use pedagogical strategies that respond to this problem. They devote most of their class time to lecture, whole-class recitation, and seatwork, activities that can be easily monitored and which keep most students busy (Goodlad, 1984). It may be that when class size drops, most teachers maintain the same instructional regime (Firestone, Mayrowetz, & Fairman, 1998). If this scenario is correct, then student achievement would typically be unaffected by class size reductions, because teachers continue teaching as they are accustomed, and the modal approach to teaching – lecture, recitation, and seatwork – works about the same whether classes are large or small.

In some cases, instructional goals require alternate approaches. Instruction of young children, in particular, typically involves substantial small-group work and individual assistance, and these may benefit from small classes. Also, the personal relationship between teachers and students may be especially important in early elementary school (Parsons, 1959; Berrueta-Clement, Schweinhart, Barnett, Epstien, &

Weikart, 1984). Finally, some teachers, regardless of grade level, favor instructional approaches that emphasize problem-solving, discussion, extensive writing, and small groups without pre-scripted activities, and students who encounter such instruction may obtain higher test scores if their classes are smaller, since this approach to teaching seems likely to be more effective in smaller classes.

What research is available to substantiate or refute these speculations? The vast majority of research on class size – both experimental and non-experimental – has only considered the number of students in the class and its relation to achievement, while neglecting to measure possible mediating conditions such as classroom instruction. A few studies have provided information on classroom conditions that may be associated with class size, and these shed some light on the question. In addition, we may use findings about where class size effects appear, and where they do not, to strengthen our confidence in different possible interpretations of how class size affects achievement.

A. Early Studies of Class Size and Instruction

In a recent manifesto, Charles Achilles argued that reducing class size fosters better teaching (Achilles, 1999). Achilles' claims reflect the conclusions of several earlier reviews (Glass, & Smith, 1980; Glass, Cahen, Smith, & Filby, 1982; Cooper, 1989), and if this conclusion were correct, the benefits of class size reduction for student learning would be self-evident. Unfortunately, the evidence supporting Achilles' claim is weak.

Research on class size and instruction goes back to a series of studies from the 1940s, 1950s, and 1960s, carried out mainly at Teachers College, Columbia University, and at the University of Texas at Austin.[27] These studies examined a variety of

classroom conditions, ranging from curriculum to instructional organization to teacher innovation and creativity, in smaller and larger classes. In some studies the size of classes varied naturally, and in others class size was deliberately manipulated. Almost all the studies found some differences that were statistically significant, but the differences tended to be small and inconsistent from study to study. The inconsistencies reflected two limitations of this research: First, there was no common conceptual framework for how class size may affect instruction, and thus no common set of practices and resources that were examined across studies. Second, measures of instruction were poorly operationalized; they were not well defined, making replication difficult, and many studies seem to have relied on subjective judgments rather than objective assessment to identify instructional differences.[28] Perhaps the only finding that appeared repeatedly in several studies was that teachers tended to provide more individualized instruction in smaller classes.

B. Observational Studies of teachers in Larger and Smaller Classes

Mary Lee Smith and Gene Glass (1980) incorporated some of the early studies into a controversial meta-analysis, which concluded, with little distinction among the various indicators of instruction, that reducing class size was associated with better teaching. In an effort to build on this work, a team of researchers at the Far West Laboratory embarked on the "Class Size and Instruction Program" (CSIP), in which they worked with one school in Virginia and one in California to reduce class size, while monitoring changes in instruction (Cahen, Filby, McCutcheon, & Kyle, 1983). In mid-year, two second-grade classrooms in each school were divided into three, reducing class size by about one-third (from 19 or 20 to 13 in Virginia, and from about 33 to 22 in

California). The investigators carried out systematic structured observations in each classroom, before and after the class size reductions. This design is especially vulnerable to a "Hawthorne effect," as the teachers were clearly aware of the purpose of the study, were selected due to their interest in trying out smaller classes, and thus would have had an artificial incentive to improve their teaching in the smaller classes. Thus, this study may shed light on what is *possible* with class size reduction rather than what is *general*. Nevertheless, the authors found that teachers' instructional approaches were substantially the same before and after the class size reduction. Discussing the Virginia case, the authors (Cahen, Filby, McCutcheon, & Kyle, 1983) noted that the teachers

> "…appeared to lack knowledge of various instructional strategies; rather, they relied on a similar format for most lessons. So, both before and after the class-size change, they most often gave brief introductions and directions for lessons rather than incorporating motivational discussions, demonstrations, or other activities. They did not provide opportunities for students to become involved in expressive activities or in small-group or individual projects."

There were no changes in instructional content or in the way content was presented. In quantitative comparisons, the only instructional differences observed were that students spent more time academically engaged, and less time off-task, in the smaller classes.

In a British study of what may be *possible* with small classes, Hargreaves, Galton, and Pell (1997, 1998) examined "expert teachers" whom they expected to adapt their teaching strategies to cope with new situations. The researchers observed 14 teachers in three settings: In their own class, in another teacher's class (which had a different number of students) and, for a subset of the teachers, in their own class with

only half the students. Thus, like the CSIP study, the researchers focused on the same teachers instructing classes of different size. On the whole, few clear differences emerged from the comparisons, although there was a tendency for more sustained interactions between teachers and students in the smaller class settings. In addition, teachers spent significantly more time on general monitoring and classroom management in larger classes, a finding that seems consistent with the CSIP observation of less off-task behavior in smaller classes.

C. Findings from Experimental Research

Using an experimental design better suited to finding generalizable results, Stan Shapson and his colleagues likewise reported modest differences between classes of varied sizes (Shapson, Wright, Eason, & Fitzgerald, 1980). Shapson examined 62 fourth- and fifth-grade classes in Toronto, in which students were randomly assigned to class sizes of 16, 23, 30, or 37 students. Classroom observations revealed virtually no consistent differences in classroom practices in smaller versus larger classes. The researchers noted that the proportion of students addressed individually by teachers was higher in smaller classes, but no more time was spent on individualization in smaller classes. In other words, the amount of individual attention was the same in different size classes, but the constant amount of time divided by smaller numbers of students allotted more time per student. Moreover, in contrast to the CSIP and the British study, Shapson found no relations between class size and student engagement or "classroom atmosphere"

Data from observations of Project STAR teachers also failed to yield a consistent pattern of instructional differences across larger and smaller classes. Even a

summer professional development institute did not lead to instructional modifications among teachers in smaller classes. Evertson and Randolph (1989, p. 102) concluded, "our findings show that teaching practices did not change substantially regardless of class type assignment or training condition."

D. Evidence from Survey Analyses

Researchers have also examined three large-scale survey data sets for correlational evidence of class size and instructional practices. These studies yield weak associations between class size and instruction, and they are not consistent with one another. In a study of 63 fifth grade classes in Australia, using both teacher questionnaires and classroom observations to collect data on instruction, Sid Bourke (1986) found more whole-class teaching, and not more individualization, in smaller classes. Smaller classes were also associated with fewer teacher-student interactions, fewer student questions, less teacher lecturing, and more probing and waiting for responses when teachers asked students questions. There was no relationship between class size and student engagement, but teachers in large classes spent more time on classroom management. However, time on classroom management was, by itself, unrelated to achievement. No conceptual basis is provided that may account for the particular pattern of associations with class size, and the study contradicts the one consistent finding from earlier studies, in which more individualization occurred in smaller classes. When Bourke combined the activities that were significantly correlated with class size into a single scale, he found the scale to be positively associated with achievement, but by eliminating non-significant elements of instruction, Bourke may have introduced biases of unknown direction.

Two U.S. national surveys relied on teacher reports to indicate instructional conditions. Analyzing the National Educational Longitudinal Survey (NELS), which began with eighth graders and their parents and teachers in 1988 with follow-ups in 1990 and 1992, Jennifer King Rice (1999) found modest associations between class size and teacher practices in high school mathematics and science. In an analysis of 3,828 science classes, she found no connections between class size and teaching. Among 4,932 mathematics classes, however, teachers in smaller classes spent more time with individuals and small groups, devoted more time to whole-class discussion, and more often engaged in practices that Rice characterized as innovative (using film, student-led discussions, small groups, and oral reports), compared with teachers in larger classes. Also in both mathematics and science, teachers in larger classes spent more time on classroom management, although this effect diminished as class size grew larger than 20 students.

In another national survey covering similar grade levels, Julian Betts and Jamie Shkolnik (1999) examined 2,170 mathematics classes from the Longitudinal Survey of American Youth (LSAY), a survey conducted of students and their parents and teachers from grades 7 through 12 in the late 1980s through the early 1990s. Betts and Shkolnik found no association between class size and text coverage and, correspondingly, no more time devoted to new material in classes of one size or another. However, smaller classes tended to spend more time on review. The authors also found teachers reporting more time on individualization in smaller classes, consistent with Rice's study and unlike the Australian elementary-school survey. Also consistent with Rice's study, Betts and Shkolnik found that teachers of large classes reported spending more time on discipline.

Overall, however, Betts and Shkolnik concluded that associations between class size and instruction in middle and high school mathematics classes are small, and probably of little practical significance. For example, their regression results imply that cutting a class size from 40 to 20 students would result in only about 3% more time devoted to review.

Correlational studies such as these are vulnerable to concerns about unobserved selectivity. The studies were carried out under the assumption that changing class size may affect what teachers do, but it could also be that teachers committed to certain practices (e.g., individualization) seek out and obtain opportunities to instruct small classes. A small class is one of the few rewards that a principal can provide to a teacher, so it is plausible that better teachers get the opportunity to teach smaller classes (although norms of equity and union contracts both militate against such favoritism). In any case it is difficult to be confident that teaching practices are endogenous to class size in survey studies.

The work of Betts and Shkolnik is especially important in this regard. Because many LSAY teachers taught more than one class in the LSAY sample, it was possible to examine how the same teacher taught classes of different size. Betts and Shkolnik used fixed-effects models to focus on within-teacher variation in the association between class size and teaching behavior. The fixed-effects results closely replicated the least-squares regressions, yielding at best modest associations between class size and teacher practices. Because of its methodological rigor, Betts' and Shkolnik's study should receive special weight in reaching conclusions about the effects of class size on teaching.

Most recently, a consortium examining the effects of class-size reduction in third grade in California reported findings on the relation between class size and instruction that were similar to the results of the national secondary school surveys (CSR Research Consortium, 1999). Generally, teacher practices in small classes (maximum 20 students) were similar to those in large classes (maximum 33). Content coverage did not differ by class size, nor did time spent on mathematics or language arts. Teachers in smaller classes spent slightly less time in whole-class instruction and more time in small-group instruction. There was no general difference in individualized instruction, although teachers in smaller classes reported giving slightly more individual help to poor readers, averaging at least five minutes of help three times per week instead of two and a half times per week as in larger classes. This finding may be another version of the greater attention to review that Betts and Shkolnik reported for secondary schools, but, as in the two national surveys, class size differences in instruction appear very small. In addition, teachers in larger classes spent slightly more time disciplining students (see further Stasz, & Stecher, 2000). This finding is also consistent with the national survey results.

Before reaching a conclusion, there is one other approach to examining the association between class size and classroom conditions that we may consider. Many researchers have asked teachers their impressions of teaching smaller classes, and these impressions are invariably favorable.[29] Teachers think that smaller classes allow more time for each individual child, that managing student behavior is easier, that student misbehavior is lessened, and that it is easier to engage students in academic activities, when there are fewer students in the classroom. However, these impressions do not correspond to anything teachers are doing differently, at least as evident from

observations and from teacher survey responses. Either teachers do not actually vary their practices in smaller and larger classes, or the instruments that have assessed classroom practices are not sensitive enough to detect the differences that do occur.

A telling result from this type of evidence is that interviews with Project STAR teachers found similar responses from teachers of smaller classes *and* teachers in regular classes with aides: both groups thought the experimental arrangement brought more time for instruction, more individualization, and fewer behavior problems (Johnston, 1989). Of course, only the smaller classes – and not those with aides – yielded higher achievement. This suggests that more individualization and fewer behavior problems may not matter or, more likely, that evidence of this sort has little value for understanding differences between what goes on in larger and smaller classes.

Overall, the weight of the evidence tilts strongly toward a conclusion that reducing class size, by itself, does not typically affect the instructional activities that occur in classrooms. We have found nothing to overturn a similar conclusion reached by Robert Slavin (1989) over ten years ago:

> "Teachers' behaviors do not vary much with the size of classes…. More accurately teachers do change their behaviors in small classes, but the changes are relatively subtle and unlikely to make important differences in student achievement."

E. Accounting for the Benefits of Small Classes in the Early Elementary Grades

If classroom instruction is more or less the same in small and large classes, how can we explain the class size effects on achievement that have been reported? In particular, how can we account for the benefits of smaller classes in experimental studies

such as STAR? The most likely explanation is that teachers whose instructional methods benefit from smaller classes – e.g., those who work with small groups, those who depend on personal relationships with students, those who emphasize hands-on projects – are more productive with smaller than with larger classes. This interpretation is consistent with the finding that class size effects occur in the early elementary grades, are substantial and persistent, but do not cumulate beyond first or second grade. Kindergarten and first grade teachers in particular are especially likely to use small groups, hands-on projects, and rely on personal relationships with students, in contrast to teachers of older children whose instruction consists largely of whole-group lecture, recitation, and seatwork.[30] Under this scenario, smaller classes would be more productive in the early grades but make little difference for achievement later on. To confirm this interpretation, we need research that assesses the interactive effects of class size and instructional activities on achievement, at a variety of grade levels. This work has not yet been done.

Existing studies of class size and instruction also give some hint about why class size effects appear larger for more disadvantaged students. Betts' and Shkolnik's analysis of national survey data indicated that teachers spend more time on review in smaller classes, and the California class size reduction study noted that third grade teachers in smaller classes spent slightly more time with poor readers compared with teachers in larger classes. If having smaller classes fosters more individual attention for students who are struggling, this may explain both the benefits for disadvantaged students, and the modest effects on average achievement in later elementary and secondary grades. The extra attention may be extremely important for students who are

falling behind, but may have only a slight impact on the overall average achievement of students in the class.

Finally, Finn and Achilles (1999) noted that students misbehave less in smaller classes, and several survey studies indicated that teachers in smaller classes spend proportionately less time on discipline (Rice, 1999; Betts, & Shkolnik, 1999; Stasz, & Stecher, 2000). Although the differences are modest, they may contribute to achievement differences, and they may help low achievers most of all. According to Gamoran and his colleagues (1995), higher rates of off-task behavior are harmful to low-achieving students but do not affect average achievement in high-achieving classes. If class size reductions stem off-task behavior, that may be especially helpful to disadvantaged students whose are over represented among those with low levels of achievement.

F. Why Does Instruction Not Vary with Class Size?

Why do most teachers follow the same instructional approach, regardless of how many students are enrolled in their classes? We can understand this finding if we recognize that schools are "institutionalized" organizations, that is, they are "*infuse with value* beyond the technical requirements of the task at hand" (Selznick, 1957, p. 17, italics in the original). Unlike businesses, which adjust their practices according to marketplace responses, schools are most sensitive to societal norms about how they should appear (Scott, & Meyer, 1983). Teaching, too, is an institutionalized practice. Teachers tend to have set ideas about what teaching means, and they follow those ideas more or less irrespective of the surrounding structural conditions. In schools, structure and activities tend to be disconnected; the way the school is organized has only a modest impact on the activities that occur inside it (Meyer, & Rowan, 1983). This has the benefit

of buffering classrooms from external pressures, but it also means that change in classrooms occurs slowly if at all.

There are also technical aspects of teaching that make it resistant to change. Teachers generally work in isolation; they close their classroom doors and generally neither observe one another nor are observed by others (Dreeben, 1970; Lortie, 1975; Johnson, 1990). Conversations among teachers generally focus on individual students, or on logistical matters, and infrequently on instructional practices (Lortie, 1975). In addition, the need to use textbooks and cover particular content places a constraint on teaching that may result in common activities regardless of class size (Cahen, Filby, McCutcheon, & Kyle, 1983). New pressures from which teachers are *not* buffered – mandated curricula and testing regimes – may further impose standardization on teaching practices irrespective of class size (Evertson, & Randolph, 1989).

The finding that teaching practices do not vary with class size is consistent with recent work on school restructuring. Observers report that teaching methods are highly resistant to changes in school structure; as Penelope Peterson, Sarah McCarthey, and Richard Elmore (1996) explained, "changing practice is primarily a problem of [teacher] learning, not a problem of organization". This does not mean teachers *cannot* change practice along with class size reductions, but it may take time, and may require opportunities for teachers to learn about other approaches to teaching. In fact some of the earliest studies of class size reduction concluded that there is a lag between the time class sizes are reduced and the time that teachers change their behavior, and the California study offered the same speculation (Richman, 1955; CSR Research Consortium, 1999).

V. Implications of the Class Size Findings

Experimental psychologists have long distinguished between the *internal validity* and *external validity* of an experiment (Campbell, & Stanley, 1966). Internal validity refers to whether one can logically infer a cause and effect relationship from an experiment or quasi-experiment that has been conducted. The external validity of an experiment refers to whether it can be generalized to other populations, other times and other scales of treatment. An experiment should have external validity before one considers basing wide spread public policy on it.

Suppose that we take at face value the findings from the Tennessee experiment that appear to indicate that class size reductions in the early grades have a long-lasting impact and that this impact is greatest for students from disadvantaged backgrounds. Our review of findings from large-scale quasi-experimental studies from other countries tends to support the Tennessee results. What are the implications of these findings for public policy? When we say we take the findings of the Tennessee experiment at face value, this means we believe the experiment had internal validity. However, there are a number of factors that lead us to question whether the external validity of the Tennessee experiment has been established sufficiently to warrant generalizing across different populations and settings in the US. Our view is that we need more and varied randomized class-size experiments, balanced with quasi-experiments that employ multi-level longitudinal designs. This lack of external validity has not prevented large-scale class size reduction initiative from being instituted by both federal and state governments in the United States. Our discussion here is meant to pose a cautionary note.

Class size reduction initiatives presuppose the availability of teachers who are equivalent in quality to existing teachers to staff the extra classrooms. Leaving aside for a moment how one might measure teacher quality, if students' learning is related to the quality of their teachers, if the teachers hired to staff the new classrooms are of lower quality than existing teachers, student learning is unlikely to increase by as much as the experimental evidence predicts it will.

Many school districts are facing great difficulty in finding qualified teachers to staff their schools and a large-scale class size reduction policy would exacerbate this problem. Certainly the evidence from California, which has implemented a class size reduction effort statewide, suggests this issue should be a serious concern of policymakers (Stecher, & Bohrnstedt, 2000). While one might instead try to provide teachers with more support in their classrooms, for example by keeping class size constant but providing the teachers with more aides, the Tennessee experiment suggests that having more teacher aides in classrooms does little to improve student learning. Efforts to increase the supply of qualified teachers are likely to require increases in teacher salaries to attract more people into the profession.

Even if qualified teachers could be found, institution of a large-scale class size reduction program presupposes the existence of vacant classrooms into which the new classes created reductions could be placed.[31] If schools were operating at or near capacity, class size reductions would require the construction of new facilities, which would add to the cost of the program. Schools were allowed to participate in Project Star only if they had sufficient excess capacity in which to place newly created classes.

Our discussion suggests that even if one were sure about what the impact of a class size reduction policy would be on students' learning, the desirability of implementing such a policy would depend on a careful weighing of its benefits and costs and of alternative policies designed to accomplish the same goal. In particular, the case is being made in many countries that increasing investments in the early years, from conception to age five, are more likely to bring greater long-term gains in children's development than increased investments at the primary or secondary school levels (McCain, & Mustard, 1999). However, rarely does the discussion accompanying class size reductions move to this level.

There is considerable debate in the educational community about the characteristics that make a teacher a "high quality" teacher. However, evidence from nonexperimental studies indicates that certain teacher characteristics do matter. As noted earlier, the evidence suggests that teachers with higher verbal ability and (at the secondary level) with greater subject matter knowledge are associated with greater student learning (Ehrenberg, & Brewer, 1995; Ferguson, 1991; Ferguson, & Ladd, 1996; Strauss, & Sawyer, 1986; Monk, 1994; Monk, & King, 1995). However, in spite of this evidence, school districts do not appear to systematically choose from their pools of teacher applicants those applicants who have the strongest academic backgrounds, who come from the better academic institutions, or who score the highest on tests of academic aptitude (Ballou, & Podgursky, 1997; Strauss, 1993).

Placing more weight on the academic aptitude and subject matter competencies of applicants in hiring decisions from existing pools of teacher applicants would be a relatively costless way of improving student learning. To increase the flow of high

aptitude college graduates into the teaching profession will likely require higher compensation for teachers. However, no comparative study of the relative costs of improving student learning through attracting higher quality teachers versus reducing class size has been undertaken.

Similarly any given expenditure on class size reductions could instead be used to increase teacher compensation in ways that potentially might improve student learning even more. Given the importance of teacher subject matter competence, consideration might be given to teacher compensation systems that provide extra compensation for enhanced subject matter knowledge. Consideration might also be given to providing financial incentives to a school's teachers, as a group, for improving their students' academic performance. Similarly, given the evidence that teacher sick leave provisions in union contracts influences teacher absenteeism, teacher absenteeism influences student absenteeism and student absenteeism influences the amount students learn, considerations might be given to financial incentives to reduce teacher absenteeism, such as allowing teachers to "cash in" their unused sick leave days when they retire or leave the district (Ehrenberg, Ehrenberg, Rees, & Ehrenberg, 1991).

In the private sector companies regularly tie their CEOs compensation to measures of company performance. However, there is evidence, at least for New York State, that school superintendents' compensation and mobility to higher paying positions in larger and wealthier districts is at best only weakly tied to the educational performance of their districts' students (Ehrenberg, Chaykowski, & Ehrenberg, 1988). Consideration might be given to tying superintendents' compensation more directly to their students'

educational gains. Given the key role of principals as educational leaders, similar incentive compensation programs might be developed for them.

Our point is that reductions in class sizes are but one of a number of policy options that can be pursued to improve student learning. Careful evaluations of the impacts of other options, preferably through the use of more true experiments, along with an analysis of the costs of each option need to be undertaken. However, to date there are relatively few studies that even compute the true costs of large class size reduction programs, let alone ask if the benefits in terms of improved student-learning merits incurring the costs.[32]

References

Achilles, C. (1999). Let's Put Kids First, Finally: Getting Class Size Right. Thousand Oaks, CA: Corwin Press.

Akerhielm, K. (1995). Does Class Size Matter? Economics of Education Review, 14, 229-241.

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's Schools: Equity at Zero Cost? Journal of Policy Analyis and Management, 15, 623-645.

Angrist, J., & Levy, V. (1999). Using Maimondides' Rule to estimate the Effect of Class Size on Scholastic Achievement. Quarterly Journal of Economics, 114, 533-575.

Ballou, D., & Podgursky, M. (1997). Teacher Pay and Teacher Quality (chapter 4). Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Barr, R. (1987). Classroom Interaction and Curricular Content. In D. Bloome (Ed.), Literacy and Schooling (pp. 150-168). Norwood, NJ: Ablex.

Barr. R. (1975). How Children Are Taught to Read: Grouping and Pacing. School Review, 84, 479-498.

Barr, R., & Dreeben, R. (1983). How Schools Work. Chicago, IL: University of Chicago Press.

Barr. R., & Sadow, M. (1989). Influence of Basal Programs on Fourth-Grade Reading Instruction. Reading Research Quarterly, 24, 44-71.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). Mathematics Achievement in the Middle School Years: IEA's

Third International Mathematics and Science Study (TIMSS). Chestnut Hill, MA: Boston College.

Berrueta-Clement, J., Schweinhart, L., Barnett, W., Epstein, A., & Weikart, D. (1984). Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19. Ypsilanti, MI: High/Scope Pulishers.

Betts, J. (1995). Does School Quality Matter: Evidence from the National Longitudinal Survey of Youth. Review of Economics and Statistics, 77, 231-250.

Betts, J. (1996). Is There a Link Between School Inputs and Earnings? Fresh Scrutiny of an Old Literature. In G. Burtless (Ed.), Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success. (141-191). Washington, DC: Brookings Institution Press.

Betts, J. R., & Shkolnik, J. L. (1999). The Behavioral Effects of Variations in Class Size: The Case of Math Teachers. Educational Evaluation and Policy Analysis, 21, 193-215.

Bishop, J. J. (1998). The Effect of Curriculum-Based External Exit Exam Systems on Student Achievement. Journal of Economic Education, 29, 171-182.

Bok, D.(1993). The Cost of Talent (pp. 57-60). New York, NY: The Free Press

Bohrnstedt, G. W., & Stecher, B. M. (Eds.) (1999). Class Size Reduction in California: Early Evaluation Findings, 1996-98. Palo Alto, CA: American Institutes for Research.

Bourke, S. (1986). How Small is Better: Some Relationships between Class Size, Teaching Practices, and Student Achievement. American Educational Research Journal, 23, 558-571.

Bowles, S., & Levin, H. (1968). The Determinants of School Achievement: An Appraisal of Some Recent Evidence, Journal of Human Resources, 3, 3-24.

Brewer, D. J., Krop, C., Gill, B. P., & Reichardt, R. (1999). Estimating the Cost of National Class Size Reductions Under Different Policy Alternatives. Education Evaluation and Policy Analysis, 21, 179-192.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods. Newbury Park, CA: Sage.

Cahen, L. S., Filby, N., McCutcheon, G., & Kyle, D. W. (1983). Class Size and Instruction. New York, NY: Longman.

Cain, G., & Watts, H. (1970). Problems in making Policy Inferences from the Coleman Report, American Sociological Review, 35, 228-242.

Campbell, R., Hombo, C. M., & Mazzeo, J. (2000). NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance. Washington, DC: U.S. Department of Education.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and Quasi-Experimental Designs for Research. Boston, MA: Houghton Mifflin.

Card, D., & Kruger, A. B. (1996). Labor Market Effects of School Quality: Theory and Evidence. In G. Burtless (Ed.), Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success (pp. 97-141). Washington, DC: Brookings Institution Press.

Card, D., & Kruger, A. B. (1992). Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. Journal of Political Economy, 100, 1-40.

Case, A., & Yogi, M. (1999). Does School Quality Matter? Returns to Education and the Characteristics of School is South Africa. National Bureau of Economic Research, Working Paper W7399.

Coleman, James S., & et. al. (1966). Equality of Educational Opportunity. Washington, DC: U. S. Government Printing Office.

Cook, T. D., & Campbell, D. T. (1979). Design and Analysis for Field Settings. Chicago, IL: Rand McNally.

Cooper, H. M. (1989). Does Reducing Student-to-Instructor Ratios Affect Achievement? Educational Psychologist, 24, 79-98.

CSR Research Consortium (1999). Class Size Reduction in Claifornia 1996-98: Early Findings Signal Promise and Concerns. Palo Alto, CA: CSR Research Consortium.

Cynader, M. S., & Frost, B. J. (1999). Mechanisms of Brain Development: Neuronal Sculpting by the Physical and Social Environment. In D. P. Keating and C. Hertzman (Eds.), Developmental Health and the Wealth of Nations (Chapter 8). New York, NY: The Guilford Press.

Dreeben, R. (1970). The Nature of Teaching: Schools and the Work of Teachers. Glenview, Il: Scott, Foresman.

Educational Research Services (1978). Class Size: A Summary of Research. Arlington, VA: Educational Research Services.

Educational Research Services (1980). Class Size Research: A Critique of Recent Meta-Analyses. Arlington, VA: Educational Research Services.

Ehrenberg, R. G., & Brewer, D. J. (1995). Did Teachers' Verbal Ability and Race Matter in the 1960s? Coleman Revisited. Economics of Education Review, 14, 1-21, 291-299.

Ehrenberg, R. G., Chaykowksi, R. P., & Ehrenberg, R. A. (1988). Determinants of the Compensation and Mobility of School Superintendents. Industrial and Labor Relations Review, 41, 386-401.

Ehrenberg, R. G., Ehrenberg, R. A., Rees, D., & Ehrenberg, E. L. (1991). School District Leave Policies, Teacher Absenteeism, and Student Achievement. Journal of Human Resources, 26, 72-105.

Ehrenberg, R. G., Goldhaber, D., & Brewer, D. J. (1995). Did Teachers Race, Gender, and Ethnicity Matter: Evidence from NEL888, Industrial and Labor Relations Review, 48, 547-561.

Evertson, C. M., & Randolph, C. H. (1989). Teaching Practices and Class Size: A NewLook at an Old Issue. Peabody Journal of Education, 67, 85-105.

Ferguson, R. F. (1991). Paying for Public Education: New Evidence on How and Why Money Matters. Harvard Journal of Legislation, 28, 465-497.

Ferguson, R. F., & Ladd, H. F. (1996). How and Why Money Matters: An Analysis of Albabama Schools. In H. F. Ladd, (Ed.), Holding Schools Accountable: Performance Based Reform in Education (pp. 265-298). Washington, DC: Brookings Institution Press.

Finn, J. P., & Achilles, C. M. (1990). Answers and Questions About Class Size: A Statewide Experiment. American Education Research Journal, 27, 557-577.

Finn, J. P., & Achilles, C. M. (1999). Tennessee's Class Size Study: Findings, Implications and Misconceptions. Educational Evaluation and Policy Analysis, 21, 97-110.

Finn, J. D., Fulton, D., Zaharias, J., & Nye, B. A. (1989). Carry-Over Effects of Small Classes. Peabody Journal of Education, 67, 75-84.

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (in press). The Enduring Effects of Small Classes. Teachers College Record.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance Based Assessment and Instructional Change: The Effect of Testing in Maine and Maryland. Educational Evaluation and Policy Analysis, 20, 95-113.

Frempong, G., & Willms, J. D. (in press). The Importance of Quality Schools. In J. D. Willms (Ed.), Vulnerable Children: Findings from Canada's National Longitudinal Survey of Children and Youth. Edmonton: University of Alberta Press.

Gamoran, A., Nystrand, M., & Berends, M., & LePore, P. C. (1995). An Organizational Analysis of the Effects of Ability Grouping, American Educational Research Journal, 32, 687-715.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1989). School Class Size: Research and Policy. Beverly Hills, CA: Sage Publications.

Glass, G. V., & Smith, M. L. (1979). Meta-Analysis of Research on Class Size and Achievement. Educational Evaluation and Policy Analysis, 1, 2-16.

Glass, G. V., & Smith, M. L. (1979). Meta-Analysis of Research on Class Size

and it Relationship to Instruction and Attitudes. American Educational Research Journal, 17, 419-433.

Goldhaber, D. D., & Brewer, D. J. (1997). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. Journal of Human Resources, 32, 505-523.

Goldstein, H., (1995). Multilevel Statistical Models (2nd Ed.). London: Arnold.

Goldstein, H., & Blatchford, P. (1998). Class Size and Educational Achievement: A Review of Methodology with Special reference to Study Design. British Educational Research Journal, 24, 255-268.

Goodlad, J. I. (1984). A Place Called School. New York, NY: McGraw-Hill.

Grissmer, D. (1999). Conclusions: Class Size Effects: Assessing the Evidence, Its Policy Implications and Future Research Agenda. Educational Evaluation and Policy Analysis, 21, 179-192.

Grissmer, D., & et. al., (1994). Student Achievement and the Changing American Family. Santa Monica, CA: Rand Corporation.

Haberman, M., & Larson, R. (1968). Would Cutting Class Size Change Instruction? The National Elementary Principal, 47, 18-19.

Hanushek, E. (1997). Assessing the Effects of School Resources on Student Performance: An Update. Educational Evaluation and Policy Analysis, 19, 141-164.

Hanushek, E. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. Journal of Economic Literature, 24, 1141-1177.

Hanushek, E. (1999). The Evidence on Class Size. In S. Mayer, & P. Peterson (Eds.), Earning and Learning: How Schools Matter (pp. 131-168). Washington, DC: Brookings Institution Press.

Hanushek, E. (1989). The Impact of Differential expenditures on School Performance. Educational Research, 18, 45-51.

Hanushek, E. (1996). School Resources and Student Performance. In G. Burtless (Ed.), Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success. (pp. 74-92). Washington, DC: Brookings Institution Press.

Hanushek, E. A. (1999). Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects. Educational Evaluation and Policy Analysis, 21, 143-164.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1999). Do Higher Salaries Buy Better Teachers? National Bureau of Economic Research, Working Paper 7082.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). Teacher, Schools and Academic Achievement. National Bureau of Economic Research. Working Paper 6691.

Hanushek, E. A., & Rivkin, S. G. (1997). Understanding the Twentieth-Century Growth in School Spending. Journal of Human Resources, Winter, 1997) 35-68.

Hargreaves, L., Galton, M., & Pell, A. (1997). The Effects of Major Changes in Class Size on Teacher-Pupil Interaction in Elementary School Classes in England: Does Research Merely Confirm the Obvious? Paper presented at the Annual Meeting of the American Educational Research Association.

Hargreaves, L., Galton, M., & Pell, A. (1998). The Effects of Changes in Class Size on Teacher-Pupil Interaction. International Journal of Educational Research, 29, 779-795.

Heckman, J., Layne-Farrar, A., & Petra. T. (1996). Does Measured School Quality Really Matter? An Examination of the Earnings-Quality Relationship. In G. Burtless (Ed.), Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success (pp. 192-289). Washington, DC: Brookings Institution Press.

Hedges, L. V., & Greenwald, R. (1996). Have Times Changed? The Relationship Between School Resources and Student Performance. In G. Burtless (Ed.), Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success (pp. 41-92). Washington, DC: Brookings Institution Press.

Hedges, L., Laine, R., & Greenwald, R. (1994). Does Money Matter: A Meta-Analysis of Studies of Differential Schooling Spending on School Performance. Educational Researcher, 23, 5-14.

Hedges, L. V., & Olkin, I. (1985). Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press.

Hoxby, C. (1998). The Effects of Class Size and Composition on Students' Achievement: New Evidence from Natural Population Variation. National Bureau of Economic Research, Working Paper 6869.

Hoxby, C. (1998). The Effects of Class Size and Composition on Students' Achievement: New Evidence from Natural Population Variation. Quarterly Journal of Economics, 115, 1287-1315.

Jackson, P. W. (1968). Life in Classrooms. New York, NY: Rinehart and Winston.

Johnson, G. E., & Stafford, F. P. (1973). Social Returns to the Quantity and Quality of Schooling. Journal of Human Resources, 8, 139-155.

Johnson, S. M. (1990). Teachers at Work. New York, NY: Basic Books.

Johnston, J. M. (1989). Teacher Perceptions of Changes in Teaching When They Have a Small Class or an Aide. Peabody Journal of Education, 67, 106-122.

Kruger, A., & Hanushek, E. (2000). The Class Size Policy Debate, Economic Policy Institute, Working Paper, 121. Washington, DC. Available: http://epinet.org

Kruger, A., & Whitmore, D. (2001). The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR. National Bureau of Economic Research, Working Paper W7656.

Kruger, A. B. (1999). Experimental Estimates of Education Production Functions. Quarterly Journal of Economics, 114, 497-532.

Lazear, E. P. (1999). Educational Production. National Bureau of Economic Research, Working Paper 7349.

Lindbloom, D. H. (1970). Class Size as It Affects Instructional Procedures and Educational Outcomes. Minneapolis, MN: Educational Research and Development Council of the Twin Cities Metropolitan Area. ERIC document no. ED 059 532.

Lortie, D. C. (1975). Schoolteacher. Chicago, IL: University of Chicago Press.

McCain, Hon. M., & Mustard, F. (1999). Reversing the Real Brain Drain: Early Years Study. Toronto, CA: Children's Secretariate.

McPherson, M. S., & Shapiro, M. O. (1998), The Student Aid Game. Princeton, NJ: Princeton University Press.

Meyer, J. W., & Rowan, B. (1983). Institutionalized Organizations: Formal Structure as Myth and Ceremony & The Structure of Educational Organizations. In J. W. Meyer, & W. R. Scott (Eds.), Organizational Environments (pp.21-44, 71-97). Beverly Hills, CA: Sage.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Ehrle, K. (1999). Evaluating the SAGE Program: A Pilot Program in Targeted Pupil-Teacher Reduction in Wisconsin. Educational Evaluation and Policy Analysis, 21, 165-178.

Monk, D. H. (1994). Subject Area Preparation of Secondary mathematics and Science Teachers and Student Achievement. Economics of Education Review, 13, 125-145.

Monk, D. H., & King, J. A. (1995). Multilevel Teacher Resource Effects on Pupil Performance in Secondary Mathematics and Science: The Case of Teacher Subject Matter Preparation. In R. G. Ehrenberg (Ed.), Choices and Consequences: Contemporary Policy Issues In Education. Ithaca, NY: ILR Press.

Mosteller, F. (1995). The Tennessee Study of Class Size in the Early School Grades: The Future of Children, vol. 5, 113-127.

Mosteller, F., & Moynihan, D. P., Ed. (1970). On Equality of Educational Opportunity. New York, NY: Random House.

Mosteller, F., Light, R., & Sachs, J. (1996). Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size. Harvard Educational Review, 66, 797-842.

Newell, C. A., (1943). Class Size and Adaptability. New York, NY: New York Bureau of Publications, Teachers College, Columbia University.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The Long-Term Effects of Small Classes: A Five-year Follow-Up of the Tennessee Class Size Experiment. Education Evaluation and Policy Analysis, 21, 127-142.

Olson, M. R. (1971). Ways to Achieve Quality in Small Classrooms: Some Definitive Answers. Phi Delta Kappan, 53, (1), 63-65.

Otto, H. J., Condon, M. L., James, E. W., Olson, W., & Weber, R. A. (1954). Class Size Factors in Elementary Schools. Austin, TX: University of Texas Press.

Parish, T., & Brewer, D. J., (1998). Class Size Reduction Policies and Rising Enrollments: Cost Implications Over the Next Decade. Palo Alto, CA: American Institutes for Research.

Parsons, T. (1959). The School Class as a Social System. Harvard Educational Review, 29, 297-328.

Peterson, P L., McCarthey, S. J., & Elmore, R. F. (1996). Learning from School Restructuring. American Educational Research Journal, 33, 149-153.

Pugh, Jr., J. B. (1965). The Performance of Teachers and Pupils in Small Classes. New York, NY: Institute of Administrative Research, Teachers College, Columbia University.

Raudenbush, S. W., & Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. Sociology of Education, 59, 1-17.

Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. Journal of Educational and Behavioral Statistics, 20, 4, 307-355.

Rees, D. I., Argus, L. M., & Brewer, D. J. (1996). Tracking in the United States: Descriptive Statistics from NELS. Economics of Education Review, 15, 83-89.

Reichart, R., & Brewer, D. (1999). Teacher Characteristics. Santa Monica, CA: Rand Corporation Report, 61-80.

Rice, J. K. (1999). The Impact of Class Size on Instructional Strategies and the Use of Time in High School Mathematics and Science Courses. Educational Evaluation and Policy Analysis, 21, 215-229.

Richman, H. (1955). Educational Practices as Affected By Class Size. New York, NY: Teachers College, Columbia University, unpublished doctoral dissertation.

Ritter, G. W., & Boruch, R. F. (1999). The Political and Institutional Origins of a Randomized Controlled Trial on Elementary and School Class Size. Educational Evaluation and Policy Analysis, 21, 111-126.

Robinson, G. E. (1990). Synthesis of Research on Effects of Class Size. Educational Leadership, 47, 80-90.

Robinson, G. E., & Wittebols (1986). Class Size Research: A Related Cluster Analysis for Decision Making. Arlington, VA: Educational Research Services.

Ross, D. H., & McKenna, B. H. (1955). Class Size: The Multi-Million Dollar Question. New York, NY: Institute of Administrative Research, Teachers College, Columbia University.

Scott, W. R., & Meyer, J. W. (1983). The Organization of Societal Sectors. In J.W. Meyer, & W. R. Scott (Eds.), Organizational Environments (pp. 129-153). Beverly Hills, CA: Sage.

Selznick, P. (1957). Leadership in Administration (p. 17). New York, NY: Harper & Row.

Shapson, S. M., Wright, E. N., Eason, G., & Fitzgerald, J. (1980). An Experimental Study of the Effects of Class Size. American Educational Research Journal, 17, 141-152.

Slavin, R. (1989). Class Size and Student Achievement: Small Effects of Small Classes, Educational Psychologist, 24, 106.

Slavin, R. E. (1989). Achievement Effects of Substantial Reductions in Class Size. In R. E. Slavin (Ed.), School and Classroom Organization. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of Research on Class Size and Its Relation to Attitudes and Instruction. American Educational Research Journal, 17, 419-433.

Stasz, C., & Stecher, B. M. (2000). Teaching Mathematics and Language Arts in Reduced Size and Non-Reduced Size Classrooms. Educational Evaluation and Policy Analysis, 22, 313-329.

Stecher, B. M., & Bohrnstedt, G. W. (2000). Class Size Reductions in California: The 1998-99 Evaluation Findings. Sacramento, CA: California Department of Education.

Strauss, R. P. (1993). Who Should Teach in Pennsylvania's Public Schools. Pittsburgh, PA: Carnegie Mellon University Center for Public Financial Management.

Strauss, R. P., & Sawyer, E. (1986). Some New Evidence on Teacher and Student Competencies. Economics of Education Review, 5, 41-48.

U.S. Bureau of the Census (1999). Statistical Abstract of the United States: 1999 (tables 83, 272, 307, 659, 760, 761). Available: http://www.census.gov Washington, DC: U.S. Government Printing Office.

U.S. Department of Education. (1999). The Digest of Education Statistics, 1999 (tables 65, 78, 108). Washington, DC: U.S. Government Printing Office.

Welch, F. (1967). Labor Market Discrimination: An Interpretation of Income Differences in the Rural South. Journal of Political Eonomy, 75, 225-240.

Whitsett, R. C. (1955). Comparing the Individualities of Large Secondary School Classes with Small Secondary School Classes through the Use of a Structured Observation Schedule. New York, NY: Teachers College, Columbia University, unpublished doctoral dissertation.

Willms, D., & Kerckhoff, A. C. (1995). The Challenge of Developing New Social Indicators. Educational Evaluation and Policy Analysis, 17, 113-131.

Willms, D., & Raudenbush, S. W. (1997). Effective Schools Research: Methodological Issues. Research, Methodology and Measurement: An International Handbook, pages.

Willms, D., & Somers, M.A. (2000). Schooling Outcomes in Latin America: report prepared for UNESCO. Fredricton, NB: Canadian Research Institute for Social Policy.

Willms, J. D., & Raudenbush, S. W. (1992). A Longitudinal Hierarchical Lindear Model for Estimating School Effects and Their Stability. Journal of Educational Measurement, 26, 209-232.

Word, E. R., et. al (1990). The State of Tennessee's Pupil/Teacher Achievement Ratio (STAR) Project. Nashville, TN: Tennessee Department of Education.

Appendix

A Dynamic, Multilevel Model for Estimating Class Size.

A. A Simple Two-Level Multilevel Model

The logic underlying multilevel regression models is that data at the lowest level (e.g., pupils nested within classrooms) are fit to a regression model separately for each of the second-level units (e.g., classes). The parameter estimates for these regression models become the dependent variables in regression models that include variables collected at the second level. The estimation technique combines the equations at both levels into a single equation, and estimates all parameters simultaneously. We will demonstrate the approach with an example pertaining to class size.

In the Third International Mathematics and Science Study (TIMSS), data were collected for students within several hundred classrooms within each country (Beaton, Mullis, Martin, Gonzalez, Kelly, & Smith, 1996). At the first level, therefore, one could regress students' mathematics achievement scores (the dependent variable) on a set of student background factors, and a dichotomous variable, *Female*, denoting whether the student was male or female. For simplicity, we use a single variable, socio-economic status (SES) to summarize the set of relevant family background factors. The model for each class would be expressed as:

(1)

$$(Mathematics)_{ij} = \beta_{0j} + \beta_{1j}(SES)_{ij} + \beta_{2j}(Female)_{ij} + \varepsilon_{ij}$$

where $(\text{Mathematics})_{ij}$, $(\text{SES})_{ij}$, and $(\text{Female})_{ij}$ are the scores on these variables for the i^{th} student in the j^{th} class. If there were 1000 classes in the study, the analysis would furnish 1000 separate

estimates of the regression "parameters", β_0, β_1, and β_2, one set for each class. It is common practice in multilevel modeling to "center" the independent variables by subtracting the grand mean of each variable from each person's score. When this is done, β_0 indicates the expected score for a group of students that was representative of all students in the sample; or in other words, it is the average level of performance for the class, after adjusting statistically for SES and sex. β_1 and β_2 (which are unaffected by the centering) indicate, respectively, the extent of inequalities among students with differing SES, and the achievement gap between females and males.

In the first instance, we are interested in the "effect" of class size on average class performance, after taking account of potentially confounding factors. Suppose that two variables, one denoting teacher experience and another describing teaching materials could capture the confounding factors. The β_{0j}'s from equation 1 would become the dependent variable in a class-level regression on class size and the confounding variables:

$$\beta_{0j} = \Phi_{00} + \Phi_{01}(Class\ Size)_j + \Phi_{02}(Teacher\ Exp)_j + \Phi_{03}(Materials)_j + \mathbf{U}_{0j} \quad (2)$$

Here Φ_{00} is the grand mean (actually the mean of the classroom means), Φ_{01} is the "effect" associated with a one-student increase in class size, and Φ_{02} and Φ_{03} are the effects associated with teacher experience and classroom materials respectively. U_{oj} is a class-level error term.

Similarly, we can write equations for SES inequalities, $\beta_{1j,}$ and the female-male achievement gap, β_{2j}:

(3) $$\beta_{1j} = \Phi_{10} + \Phi_{11}(Class\ Size)_j + \Phi_{12}(Teacher\ Exp)_j + \Phi_{13}(Materials)_j + \mathbf{U}_{1j}$$

(4) $$\beta_{2j} = \Phi_{20} + \Phi_{21}(Class\ Size)_j + \Phi_{22}(Teacher\ Exp)_j + \Phi_{23}(Materials)_j + \mathbf{U}_{2j}$$

In equation 3, Φ_{11} is the cross-level interaction term for SES and class size, which indicates whether class size has differing effects for low or high SES students. If the estimate of the coefficient were positive and statistically significant, then we would conclude that class size had larger effects for high SES students than for low SES students. The reverse would hold if the coefficient were negative. Similarly, Φ_{21} is the cross-level interaction term for sex and class size, which indicates whether class size has differing effects for males and females. Equations 2, 3, and 4 could include different sets of covariates corresponding to one's theory about the relationships. Also, they could include same-level interaction terms, such as a "class-size-by-teacher-experience" interaction. The statistical and computing techniques upon which multilevel models are based incorporate equations 1 through 4 into a single model, and estimate the parameters of the model using iterative procedures (Raudenbush, & Bryk, 1986; Goldstein, 1995).

B. Three examples.

Willms and Kerckhoff (1995) estimated the effects of pupil-teacher ratio on the reading and mathematics scores of 16-year old pupils in 148 local education authorities in England and Wales. Their set of controls at the first level (equation 1) included sex, SES, and a prior (age 12) measure of achievement. They found a negative relationship with class size: -.018 for reading (not statistically significant), and -.032 for mathematics (significant, $p < .05$). These results suggest that a decrease in pupil-teacher-ratio from 25 to 16 would result in an increase in achievement by about .16 of a standard deviation for reading, and .29 of a standard deviation for mathematics.

Frempong and Willms (2001) estimated the effect on mathematics achievement associated with pupil-teacher-ratio for Canadian grade 7 and 8 students, using the TIMSS data. Their model included a much more comprehensive set of controls at the second level, including variables describing classroom-teaching processes. Their estimate of the class size effect was .02, which corresponds to an increase in student achievement of about .18 of a standard deviation for a decrease in pupil-teacher ratio from 25 to 16. They also detected a significant cross-level interaction with SES: low SES students were more likely to benefit from a decrease in pupil-teacher-ratio. Finally, a study of primary schooling outcomes in eleven Latin American countries yielded average estimates of the effect of pupil-teacher ratio that correspond to effect sizes of about .07 for language and .08 for mathematics (Eillms, & Somers, 2000).

All three of these studies suffer from the problems of quasi-experimentation discussed above, even though care was taken to control for students' family background and potentially confounding variables at the classroom and school levels. Also, the studies employed a measure

of pupil-teacher-ratio, instead of class size *per se*. We suspect that the effects of pupil-teacher ratio are likely to be weaker than those of class size. Nevertheless, the findings are reasonably consistent across the studies, and suggest that a decrease in pupil-teacher ratio from 25 to 16 is likely to increase student achievement by about .20 of a standard deviation in high-income countries, and by about half that in low-income countries.

C. Stronger Models for True and Quasi-Experiments: **Growth and Stability Models**

One of the chief problems with the studies above is that they are based on a single observation on one occasion. Thus, they are a measure of the students' achievement *status*, rather than their *rate of learning.* When data are collected on the same students on at least three occasions, it is possible to estimate a growth trajectory for each child, and the average growth trajectory for a class or school (Bryk, & Raudenbush, 1992). In the framework of hierarchical or multilevel models, this is accomplished with a three-level model. The first level of the model is intra-individual; it models each child's initial status and rate of learning. These parameters are carried to the next two levels, which are identical to those specified in equations 1 through 4 above. In the case of a class size experiment, the interest is in whether children's rate of growth is faster in smaller classes than in larger classes. This approach provides much more reliable estimates of effects than cross-sectional (post-test only) studies, or pre-post studies. It can be used for both true and quasi-experiments. In the case of quasi-experiments, the same threats to validity are germane; however, selection bias is less of an issue because the focus is on the rate of learning rather than students' status at some point in their educational career.

Most state Departments of Education collect statewide data on student achievement to monitor student performance. Many school districts also conduct their own performance

assessments. These monitoring systems generally furnish data that describe successive cohorts of children passing through the same classrooms and schools on an annual or biennial basis. Willms and Raudenbush (1992) set out a multilevel model for assessing the "stability" of school effects. It asks whether schools are improving or getting worse in their school performance from year to year, and assesses whether *changes* in school performance are related to *changes* in school policy and practice. The model is similar to the two-level model described above, except that a third level is included which models the changes over time of the background-adjusted estimates of school performance (i.e., the b_{0j} in equation 1). This is a potentially powerful approach to assessing the effects of a statewide policy to reduce class size, because most of the potentially confounding variables are being held constant, thereby allowing the analyst to assess whether the benefits are being realized as implementation of the intervention is proceeding apace.

Table 1

PERCENTAGE DISTRIBUTION OF ESTIMATED INFLUENCE OF PUPIL/TEACHER RATIOS ON STUDENT PERFORMANCE

		Statistically Significant		Statistically Insignficant		
	Number of Estimates	Positive	Negative	Positive	Negative	Unknown Sign
A) SCHOOL LEVEL						
All	277	15%	13%	27%	25%	20%
Elementary	136	13%	20%	25%	20%	23%
Secondary	141	17%	7%	28%	31%	17%
B) LEVEL OF AGGREGATION OF THE RESOURCE MEASURE						
Total	277	15%	13%	27%	25%	20%
Classroom	77	12%	8%	18%	26%	36%
School	128	10%	17%	26%	28%	19%
District	56	21%	16%	39%	20%	4%
County	5	0%	0%	40%	40%	20%
State	11	64%	0%	27%	9%	0%
C) VALUE ADDED MODELS (GAIN SCORE EQUATIONS)						
All Studies	78	12%	8%	21%	26%	35%
Within State Studies	23	4%	13%	30%	39%	13%

Source: Eric Hanushek, "The Evidence on Class Size" in Susan E. Mayer and Paul Peterson eds. *Earning and Learning: How Schools Matter* (Washington DC: Brookings Institution Press, 1999), Tables 4, 5 and 6. In this table positive means that smaller classes improve student performance

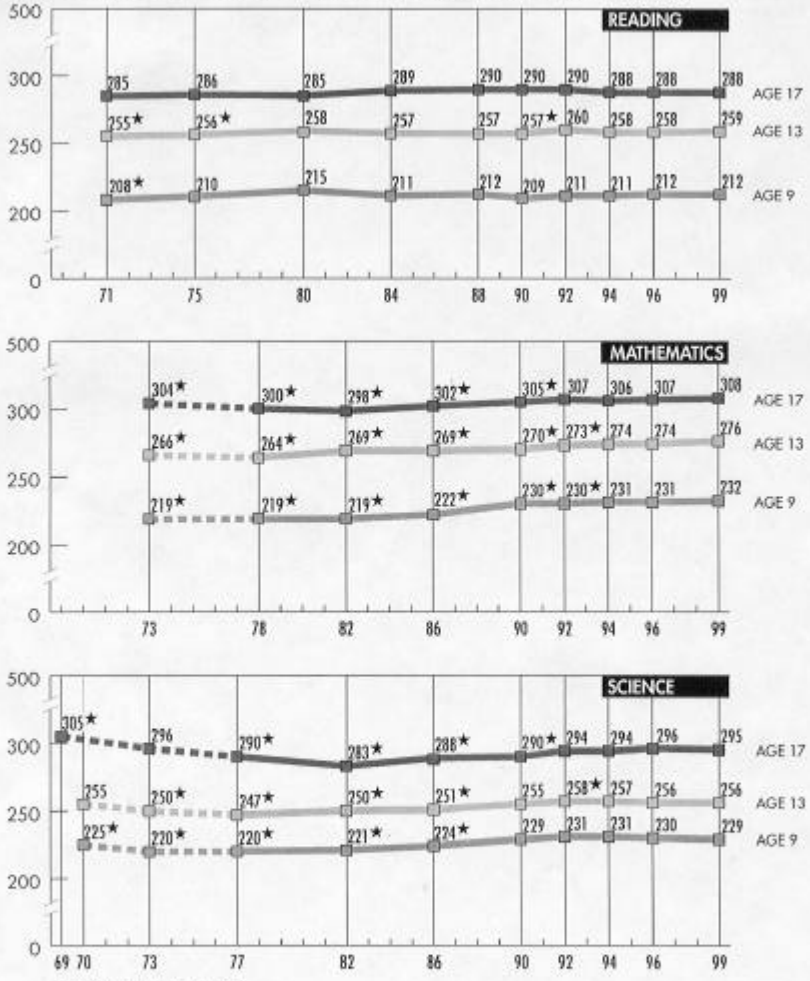
Table 2

Small Class Size Effects in Tennessee Star Study

		GRADE LEVEL							
		K	1	2	3	4	5	6	7
Total Reading	White	N/A	.17	.13	.17	N/A	N/A	N/A	N/A
	Minority	N/A	.37	.33	.40	N/A	N/A	N/A	N/A
	All	.18	.24	.23	.26	.13	.22	.21	.15
Total Mathematics	White	.17	.22	.12	.16	N/A	N/A	N/A	N/A
	Minority	.08	.31	.35	.30	N/A	N/A	N/A	N/A
	All	.15	.27	.20	.23	.12	.18	.16	.14
		5738	6752	5148	4744	4230	4649	4333	4944

Source: Finn and Achilles (1999), Tables 1 and 2.
N/A = not available

Figure 1
Trends in Average Scale Scores for the Nation in Reading, Mathematics, and Science



★ Significantly different from 1999.
NOTE: Dashed lines represent extrapolated data.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1999 Long-Term Trend Assessment.

[1] Parish and Brewer (1998), present a detailed summary of the differences in current state class size reduction policies and proposed federal ones.
[2] NAEP is a congressionally mandated and is conducted by the National Center for Education Statistics. For details of its recent sample frame, subject area coverage and results, interested readers can consult Campbell, Hombo, & Mazzeo (2000).
[3] McPherson, & Shapiro (1998), summarize the changing role of financial aid policies on college enrollment rates.
[4] *NAEP* 1999 Trends in Academic Performance, figures 2.3 and 2.4. However, in recent years there is evidence of a widening performance gap for some age students on some tests.
[5] Data on the average earnings of full-year female employees with at least 5 years of post secondary education is available from the U.S Census Bureau and is derived from the annual May Current Population Survey reports. These data are available on the World Wide Web at http://www.census.gov and can be found in tables P32, P33, P34 and P35 of the Historical Income Tables – People.
[6] Ehrenberg, & Brewer (1995) (pp.1-21) and Ehrenberg, Goldhaber, & Brewer (1995) (pp. 547-561) provide evidence that this view is not always correct.
[7] Gamoran, Nystrand, Berends, & LePore (1995) (pp. 687-715) and Argys, Rees, & Brewer (1996) (pp. 623-645) present empirical analyses of the effects of ability grouping on secondary school students.
[8] To give the reader a sense of the variables collected and included in their models, school characteristics variables included per pupil expenditures on staff, volumes per student in the library, students per teacher the presence of science lab facilities, the use of tracking, and school location. Student body and family characteristics (in addition to the racial/ethnic/gender distribution of the student body) included proportion of families that owned encyclopedias, attendance rates, parents' education and occupation levels, parental income levels, and the number and genders of parents in the home. Teachers' characteristics included experience, education levels, verbal aptitude test scores, their racial distribution, and their preferences to teach middle class students.
[9] See for example, Bowles, & Levin (1968) (pp. 3-24); Cain, & Watts (1970) (pp. 228-242); and Mosteller, & Moynihan (Ed) (1970).
[10] See for example, Hanushek (1986) (pp. 1141-1177); Hanushek (1989) (pp. 45-51); Hedges, Laine, & Greenwald (1994) (pp. 5-14); Hanushek (1997); Hanushek (1996) in Gary Burtless (Ed.); (pp.43-74); Hedges, & Greenwald (1996) in Burtless (Ed.) (pp. 74-92); and Hanushek (1999) in Mayer and Peterson (Eds) (pp. 131-165).
[11] Bishop (1998) (pp. 171-182) shows that curriculum-based exit exams in New York State (the Regents exams) that all high school students in the state must take to graduate lead New York State students to score higher on the SAT exam than students from other states, other factors held constant.
[12] Recently Alan Kruger has criticized Hanushek's approach because it weighted more heavily in the "count statistics" studies that contained more than one class size estimate. Kruger argues that such studies are more likely to be based on small data sets and thus more likely not to yield statistically significant estimates. Hanushek has offered a spirited rebuttal to this argument. We judge their debate to be inclusive. Interested readers can find this debate in "The Class Size Policy Debate", *Economic Policy Institute Working Paper 121* (Washington, DC: October 2000), which is available on the World Wide Web at www.epinet.org.
[13] These tests are described in Hedges , & Olkin (1985).
[14] See Hedges, & Greenwald (1996) (pp. 81-87). The formal statistical test that they used was the inverse chi-square (Fisher) method.
[15] For a formal model of why optimal class size should vary with student "ability", see Lazear (1999). Rees, Argus, & Brewer (1996), found that students in grades 8 and 10 in low track classes experienced smaller class sizes that their classmates in high track classes in the same grades.
[16] See Hoxby (1998). Similarly, Angrist, & Levy (1999), exploit variations in class size that are due to natural causes. They exploit the natural variation in class sizes that occurs in Israeli schools because of maximum class size limits. If a grade population is randomly high in a school in a year and the maximum size class limit is exceeded in the grade, an extra class must be created, which in turn reduces the class size for all students in the grade. Finally, Case, & Yogi (1999), exploit variations in pupil/teacher ratios that occurred due to the limited residential choice of black South Africans under the Apartheid system.
[17] Studies that have done this include Akerhielm (1995) (pp. 229-241). Akerhielm creates an instrumental variable for a student's actual class size by regressing a student's actual class size in a grade on the average class size in the student's grade in the school and the number of students in the grade in the school.

[18] Three examples are Ferguson (1991) (pp. 465-497); Ehrenberg, & Brewer (1995) (pp. 291-299); and Ferguson, & Ladd (1996) (pp. 265-298). However, not all researchers find this relationship. For example, Hanushek, Kain, & Rivkin (1999) find only weak evidence that students' achievement in Texas school districts is related to their teachers' scores on certification tests.

[19] Goldhaber , & Brewer (1997), (pp. 505-523) use NELS data to show the importance of teacher subject matter competencies in explaining tenth grade students' mathematics' test scores.

[20] See Card, & Kruger (1992) (pp. 1-40). Their work built on earlier work including Welch (1967) (pp. 225-240) and Johnson, & Stafford (1973) (pp. 139-155).

[21] Many of the details of the STAR design can be found in Word, et. al. (1990); Finn, & Achilles (1990) (pp. 557-577); Nye, Hedges, & Konstantopoulos (1999) (pp. 127-142) and the citations therein. For a discussion of the political origins of STAR see Ritter, & Boruch (1999) (pp. 111-126).

[22] See Krueger (1999) (pp. 497-532).

[23] Finn and Achilles (1999) argue that these figures are likely an underestimate due to the tendency of some of the smaller classes to drift above the range defined as small and for some regular classes to drift downwards in number due to student mobility.

[24] See Glass, & Smith (1979) (pp. 2-16); Robinson (1990) (pp. 80-90); Mosteller, Light & Sachs (1996) (pp. 797-842); and Slavin, (1989) in Slavin (Ed.) (pp. 80-90).

[25] The issue of linearity is particularly important because the costs associated with reducing classes to different levels is generally NOT linear – it is proportionately much more expensive to reduce class sizes to ever smaller levels.

[26] CSRP is a voluntary policy, although in practice almost every district has adopted it. It's funding formula, however, greatly exacerbates the inequalities among districts since it provides for a flat $ amount per student who is in a class of 20 or fewer. Consequently districts who were already operating at a class size of 20 received a windfall payment 20 times $650 per student (now $850) but had to do nothing, while those above this level still only received 20 times $650 per student if and only if they hired the extra teachers and facilities necessary to create new reduced sizes classes.

[27] Among these studies are Newell (1943); Otto, Condon, James, Olson, & Weber (1954); Richman (1955); Whitsitt (1955); Ross, & McKenna (1955); Pugh, Jr. (1965); Haberman, & Larson (1968), (pp. 18-19); Olson (1971) (pp. 63-65).

[28] These concerns were articulated beginning more than 20 years ago by Educational Research Services (1978); Educational Research Services (1980); and Robinson, & Wittebols (1986).

[29] See Richman (1955); Cahen, Filby, McCutcheon, & Kyle (1983); Shapson, Wright, Eason ,& Fitzgerald (1980); Achilles (1999); Lindbloom (1970); Johnston (1989); Molnar, Smith, Zahorik, Palmer, Halbach, & Ehrle (1999) (pp. 165-177).

[30] For contrasts between early and later elementary instruction and between elementary and secondary instruction, see the work of Rebecca Barr on literacy instruction: Barr (1975) (pp. 479-498); Barr, & Dreeben (1983); Barr, & Sadow (1989) (pp. 44-71); and Barr (1987) in Bloome (Ed.), (pp. 150-168).

[31] However, we remind the reader that class size reductions can also be accomplished by having two teachers in a classroom, as was done in Wisconsin's SAGE pilot project.

[32] Brewer, Krop, Gill, & Reichardt (1999), (pp. 179-192) provides estimates of the operational costs of a nationwide class size reduction program under various assumptions of the program's scope (all students vs. students-at-risk, the grade levels included, the maximum class size specified, whether the latter would be for each classroom, an average by school or an average by district etc.). They stress that their estimates are only for operational costs, ignore possible construction costs for needed new facilities and ignore the increases in teacher compensation that might be needed to attract the additional teachers that would be required into the profession.