CLASS SIZE AND STUDENT OUTCOMES:
RESEARCH AND POLICY IMPLICATIONS

Matthew M. Chingos

The Brown Center on Education Policy
The Brookings Institution

Published in: *Journal of Policy Analysis and Management* 32(2): 411–438 (2013)

INTRODUCTION

Schools across the U.S. are facing budgetary pressures on a scale not seen in generations. A year after the end of federal stimulus funding and with economic growth at low rates, 31 states were projecting a combined $55 billion in shortfalls for their 2013 budget year. These numbers are large by historical standards, but are dwarfed by the combined $538 billion in shortfalls that states had to close in the previous four fiscal years—an average of $135 billion per year, much if not most of which took the form of deep spending cuts. Some, including large states such as California and Texas, projected revenue shortfalls of more than 15 percent of the size of their 2013 budgets (Oliff, Mai, & Palacios, 2012). Cuts in state spending coupled with declines in property values mean that the increases in education spending that used to occur so regularly appear to have come to an end for the foreseeable future.

Times of fiscal exigency force policymakers and education practitioners to pay more attention to the return on various categories of public investment in education. The sizes of the classes in which students are educated are often a focus of these discussions because they are a key determinant of educational spending. Personnel costs constitute the single largest category of educational expenditures. The most recent data (from 2008-09) collected by the U.S.

Department of Education indicate that salaries and benefits of instructional personnel account for 45 percent of total school spending, and 91 percent of instructional expenditures.[1]

And education in the U.S. is a very labor-intensive industry. In 2010-11, our nation's public schools employed one teacher for every 16 students, and one employee of any type for every eight students (Keaton, 2012). Hiring fewer teachers is the only way to reduce instructional personnel costs without cutting salaries or benefits (which are often protected in the short term by collective bargaining agreements), so declining resources frequently lead to increases in class size—to the chagrin of parents and teachers.[2]

The declines in funding currently faced by many schools mean that cuts must be made, but it is often unclear how to make cuts in ways that minimize harm to students. For example, state lawmakers may be unsure as to the costs and benefits of maintaining a statewide cap on class size relative to other state-mandated uses of funds for education, or the likely effect of relaxing class-size mandates. Leaders of school districts may not be sure how to find the best combination of teacher compensation and class size for a given level of funding. Is it better to hire more teachers and pay them less or to hire fewer teachers and pay them more?

In this paper I review the evidence base available to inform such policy decisions. The number of high-quality studies is disappointingly small, and does not offer guidance as to the optimal class size overall, much less for specific contexts such as grades, subjects, or student populations. But it does offer lessons relevant to current policy debates. Most studies find at least some evidence of positive effects of smaller classes, but the size of these benefits is inconsistent across studies and often small. The significant costs of reducing class size coupled

[1] These figures are from the author's calculations using the NCES Common Core of Data's Local Education Agency (School District) Finance Survey (F-33) Data for 2008-09.

[2] Teacher contracts that specify an annual increase in the salary schedule imply that even constant school spending will force schools to increase class size due to the automatic increase in personnel costs.

with these modest benefits implies that many school systems in the U.S. have overinvested in class-size reduction and that increasing class size in some situations may represent a budget-cutting strategy that minimizes harm to students.

A CENTURY OF CLASS-SIZE DEBATES

Research and policy discussions about the optimal class size in our nation's schools have existed at least as long as there has been a system of universal public education. The significant expansion of access to education in the first three decades of the 20th century—when enrollment increased from 52 to 72 percent of children aged 5 to 19—forced schools to either hire more teachers or increase class size. Concerns that larger classes might harm student achievement prompted an early generation of research on class size that peaked in the 1920s, but subsided once falling birth rates reduced school enrollments and the associated upward pressure on class size (Rockoff, 2009).

Times of financial pressure on schools, due to either increased enrollment or decreased funding, make it difficult for schools to maintain their existing class sizes. But when such financial exigencies are relaxed, they are replaced by political pressure to reduce class size. Class-size reduction has the support of a broad political coalition due to its enormous popularity with parents, teachers, and the public in general. A 2007 poll of the American public found that 77 percent of respondents thought that additional educational dollars should be spent on smaller classes rather than higher teacher salaries.[3]

Common sense—correct or not—suggests that students will learn more in smaller classes because of increased opportunities to receive individualized instruction from the teacher. It is

[3] Education Next-Program on Education Policy and Governance 2007 Survey; results available at http://educationnext.org/files/EN-PEPG_Complete_Polling_Results.pdf.

not surprising that many parents prefer smaller classes based on this sort of intuition. Teachers also support smaller classes, perhaps because they find them easier to manage. The same 2007 survey found that fully 81 percent of public school employees preferred an improvement in a working condition—class size—than an increase in salary.[4] For the public as a whole, the pupil-teacher ratio represents an easy statistic to monitor as a measure of educational quality, especially before test-score data became widely available in the last decade.

As long as tax revenues allowed, lawmakers were happy to capitalize on this political support by enacting class-size reduction policies. In recent decades, at least 24 states have mandated or incentivized class-size limits in their public schools (Education Commission of the States, 2005). Class-size legislation at the state level was particularly popular in the wake of the Tennessee Student-Teacher Achievement Ratio (STAR) experiment in the late 1980s, which found that a large reduction in class size in the early grades boosted student achievement (Krueger, 1999). For example, California allocated more than $1 billion per year in the late 1990s to reduce class size in the early grades from 30 to 20 (Jepsen and Rivkin, 2009). Florida went a step further and mandated maximum class sizes in all grades through a constitutional amendment, at a cost of about $22 billion through 2011-12.[5]

The federal government also has its own program, which provided $1.2 to $1.6 billion per year from 1999 to 2001 for class-size reduction, with the goal of reducing class size in grades K–3 to an average of 18 students per class. The funds were provided to states, which distributed them to districts using a formula that incorporated poverty and enrollment data. Districts were required to spend most of the funds on teacher salaries and recruitment and training of new

[4] However, when Washington state teachers were asked in 2006 whether they preferred a $5,000 salary increase or a two-student reduction in class size (which cost roughly the same amount), fully 83 percent said they preferred the higher salary (Goldhaber, DeArmond, & DeBurgomaster, 2010).

[5] Florida Department of Education, Class Size Reduction Amendment web site, http://www.fldoe.org/classsize/.

teachers. This program was absorbed into Title II of the No Child Left Behind Act in 2001 (Millsap et al., 2004). As a result, districts are free to spend their Title II funds to reduce class size but are no longer required to allocate a specific pot of money toward this goal.

The enthusiasm of policymakers for class-size reduction policies surely reflects broad public support for these policies, but it also likely stems, at least in part, from the scarcity of variables in American K–12 education that are both thought to influence student learning and are subject to legislative action. It is straightforward for a state legislature to pass a law allocating funds for class-size reduction, and then ensuring that the money is actually spent on reducing the number of students per class by hiring more teachers. There are certainly other educational policies that states can adopt, such as teacher evaluation systems, changes to the length of the school day and year, and testing and accountability. But whereas those policies can be challenging for state lawmakers to implement and raise questions about state versus local control, statewide class-size policies often involve little more than the legislature sending checks to districts.

These policies reinforced existing trends in local school districts that date back to the early 1900s. Figure 1 shows the pupil-teacher ratio in the U.S. since the beginning of federal data collection efforts in the 1869-70 school year. The number of students per teacher has declined with few interruptions since 1898. The largest increase occurred during the Great Depression, when the pupil-teacher ratio increased by 1.1 students between 1933 and 1935. In recent decades, economic recessions had not caused the ratio to increase by more than 0.1 students per year until the 0.6-student increase between 2009 and 2010. Over the 50-year period from 1960 to 2010, the pupil-teacher ratio fell by 39 percent, or about 0.2 students per year.

The pupil-teacher ratio is not equivalent to average class size, and is nearly always smaller than average class size. In 2007-08, the pupil-teacher ratio was 15.3, but the federally administered Schools and Staffing Survey found average class sizes of 20.3 in self-contained elementary classes and 23.3 in subject-area high school classes (Coopersmith, 2009). A simple example illustrates why this is the case. A school with 10 self-contained classes of 25 students each and no other teachers would have an average class size of 25. But if the school also employs two full-time teachers to cover subjects such as art, music, and science, then the pupil-ratio ratio would be 20.8 (250 students divided by 12 teachers). Incorporating two teachers who provide pull-out instruction to students with learning disabilities into the example further reduces the ratio to 17.9.

Despite its limitations, the pupil-teacher ratio is the best proxy for class size that has been recorded for the entire country over a reasonably long period of time. The steady downward trend partly reflects an increase in educational services to students with disabilities, as required by the federal Education for All Handicapped Children Act of 1975 and its successor, the Individuals with Disabilities Education Act (IDEA) of 1990. But Hanushek (1999) presents evidence that increases in special education account for at most one-third of the fall in the pupil-teacher ratio between 1980 and 1990.[6] Specifically, he simulates what the pupil-teacher ratio would have been in 1990 had the special education pupil-teacher ratio and the share of disabled students remained at their 1980 levels.

Hanushek's finding is reinforced by national survey data that have been gathered every five years since 1960 by the National Education Association. Figure 1 shows that the class sizes reported by elementary school teachers in self-contained classrooms fell by 24 percent between 1960 and 2005, during which time the pupil-teacher ratio fell by 41 percent. Over this time

[6] See also Hanushek and Rivkin (1997).

period the gap between average class size and the pupil-teacher ratio grew by about three students. Data from high-school teachers delivering subject-area instruction do not follow the same trend and jump around to a greater extent than the elementary data, perhaps due to noise in the self-reported survey data.

The available data leave no doubt that reductions in class size have occurred with few interruptions for decades. But do students learn more in smaller classes than they do in larger classes? How large are class-size effects, and how consistent are they across studies?

RESEARCH ON CLASS SIZE

There is a large body of research on the relationship between class size and student learning. A 1979 systematic review of the literature identified 80 studies (Glass & Smith, 1979). There are surely many more today. The vast majority of these studies simply examine the association between variation in class size and student achievement. The primary difficulty in interpreting this research is that schools with different class sizes likely differ in many other, difficult-to-observe ways. For example, more affluent schools are more likely to have the resources needed to provide smaller classes, which would create the illusion that smaller classes are better when in fact family characteristics were the real reason.[7] Alternatively, a school that serves many students with behavior problems may find it easier to manage these students in smaller classes. A comparison of such schools to other schools might give the appearance that small classes produce less learning when in fact the behavior problems were the main factor.

Studies that do not carefully isolate the causal effect of class size (and only class size) produce widely varying results. Hanushek (2003) compiled 276 estimates of class-size effects

[7] West and Woessmann (2006) study data from 18 countries and find that the U.S. is the only country in which between-school sorting is regressive in that, on average, weaker students are sorted into larger classes.

from 59 studies, and found that only 11 percent of these estimates indicated positive effects of smaller classes. A similar number (9 percent) were negative, with the remaining 80 percent not statistically distinguishable from zero. Krueger (2003) argued that each study (rather than each estimate) should be given equal weight, but using this method of counting only increased the proportion of studies showing positive effects to 26 percent, with the majority showing either negative or insignificant effects. One way to interpret these tallies is that class size matters in some circumstances but not others. That may well be true, but a more likely explanation is that unreliable studies produce uneven results.

The only way to credibly measure the causal effect of class size is to compare students who are in larger or smaller classes for reasons unrelated to their achievement. This is most clearly the case in a well-executed randomized experiment, in which students and teachers are randomly assigned to smaller or larger classes. Unfortunately, in the last 75 years only one study of this type has been carried out at any significant scale.

Natural experiments, also called quasi-experiments, often provide next-best opportunities to estimate class-size effects. In these cases it is possible to identify situations in which class size changed for reasons that were plausibly unrelated to student achievement. For example, a change in class size policy that occurred at a discrete point in time might allow for a before-and-after analysis of its effects. Or an instrumental variable might be identified that affects class size but has no direct impact on student achievement.

I limit this review to the relatively small number of studies that fall into these two categories because experimental and quasi-experimental methods are most likely to yield unbiased estimates of causal effects. Purely observational studies—for example, a regression analysis of school-level data in which class size is used to predict test scores conditional on some control variables—rest on the same main assumption: that all confounding variables have been measured and controlled for. This assumption is unlikely to be true and it is usually difficult to gauge the extent to which this important shortcoming will bias the results. Quasi-experimental studies are certainly not immune from this sort of bias, but they rest on explicit assumptions that are discussed by the authors and can be evaluated by readers for their plausibility.[8]

Below I summarize the key findings of these studies and review the strengths and weaknesses of the methods that they employ to estimate the causal effect of class size (or a class-size policy). I do not use formal meta-analytic techniques because the high-quality studies of class size are too diverse for a simple summary of the estimates to be useful. A meta-analysis is a potentially useful way to summarize several studies of roughly the same intervention, such as a five-student reduction in class size financed by the state legislature. Combining estimates from several studies reduces the uncertainty inherent in the results of any individual study. But the studies of class size I review range from an experiment in which extra teachers were financed by the state to studies of naturally occurring variation in class size to evaluations of statewide policies that may have impacted the quality of the teaching workforce. Consequently, these studies are best understood by considering them one at a time, making note of their strengths, weaknesses, and relevance to different types of policy decisions.

I divide my review of the high-quality evidence on class size into three sections. First, I discuss the Tennessee STAR experiment, which is the most important and influential study because it is the only modern randomized experiment conducted at a significant scale. Second, I review the quasi-experimental evidence based on naturally occurring variation in class size that

[8] An oft-cited quasi-experimental study that I do not include is the evaluation of the Wisconsin Student Achievement Guarantee in Education (SAGE) program (Molnar et al., 1999). The two primary reasons I do not include this study are: 1) a small number of schools were included in the evaluation (30 SAGE schools and 14–17 comparison schools), and 2) the regression analysis does not adjust the standard errors for clustering by school, so it is unclear whether the estimated SAGE effects would still be statistically significant if the correct standard errors were used.

is credibly exogenous to student achievement. Finally, I review the quasi-experimental evaluations of two statewide class-size reduction policies. I examine these studies separately because in addition to offering evidence about class size they also raise important issues related to the design and implementation of class-size policies.

TENNESSEE STAR EXPERIMENT

Randomized experiments aimed at measuring the effect of class size were fairly common in the first half of the 20th century. Rockoff (2009) summarizes 24 such experiments that were conducted between 1920 and 1940; he argues that these studies were carefully designed in general. Most of these studies examined high schools, and many were carried out on a fairly small scale. Only two of the 24 studies found increased achievement in smaller classes, and several found a large-class advantage. These early studies represent interesting historical context but are of questionable relevance to current education policy because of the significant changes that have occurred in U.S. schools and wider society in the ensuing decades—including multiple wars, the end of state-sanctioned racial segregation, and increased educational opportunities for women.

The only modern randomized experiment measuring the effects of class size in U.S. schools at a significant scale is the Student-Teacher Achievement Ratio experiment, or Project STAR, which was conducted in Tennessee during the late 1980s as the result of a legislative compromise between policymakers who wanted to reduce class size across the state and their colleagues who were skeptical that it was worth the substantial cost (Ritter & Boruch 1999). Beginning with the entering kindergarten class in 1985, students and teachers were randomly assigned to a small class, with an average of 15 students, or a regular class, with an average of 23 students.[9] Thus the reduction in class size of about 8 students, or 35 percent, was quite large. Study participants over the course of the four-year experiment included 11,600 students from 80 schools (Krueger, 1999).

There are many studies based on data from the STAR experiment that take advantage of the random assignment of students and teachers to classrooms, including several important studies that are not about class size (see, e.g., Dee, 2004 and Nye, Konstantopoulos, & Hedges, 2004). The earliest papers reporting results from the STAR experiment focused on comparisons of mean student achievement in the different treatment groups (see, e.g., Finn & Achilles, 1990; Folger & Breda, 1989; and Word et al., 1990). I review Krueger's (1999) analysis of the initial test-score data because it addresses deviations from the ideal experimental design, such as non-compliance with random assignment and attrition from the data. I also summarize two recent studies that examine longer term outcomes.

Krueger's (1999) analysis of the Tennessee STAR experiment finds that elementary school students randomly assigned to small classes outperformed their classmates who were assigned to regular classes on standardized tests by about 0.22 standard deviations after four years. This effect was concentrated in the first year that students participated in the program: the small class effect in the first year was 0.12 standard deviations, with an increment of 0.035 standard deviations in each of the following years.[10] In addition, the estimated effects of class size were largest for black students, economically disadvantaged students, inner-city students, and boys.[11]

[9] Author's calculations from Project STAR data, averaged across all four years of the experiment.

[10] These statistics are calculated by converting the estimates in Table IX of Krueger (1999) to standard deviation units.

[11] Krueger (1999) reports point estimates for these subgroups of students, but does not report whether the estimated effects are statistically significantly different across subgroups.

Project STAR also randomized regular-class students between classes with and without a full-time teacher's aide and found that having an aide had no effect on test scores. In other words, reducing the classroom's student to adult ratio by adding a full-time aide had no effect on student achievement. This important finding regarding school staffing practices is often overlooked, as evidenced by the fact that the number of school aides per student increased dramatically during the two decades following the STAR experiment. Between 1992-93 and 2009-10, the number of instructional aides per student increased by 50 percent, whereas the number of teachers per student only increased by 13 percent.[12] As in the case of the overall pupil-teacher ratio, these trends likely reflect both increases in services provided to special education students (where aides play a particularly important role) as well as more general increases that affect all students.

As in most real-world experiments, there were several deviations from the ideal experimental research design in Project STAR. Data were not gathered on the class type to which students were randomly assigned (only on the class they ultimate attended), and about 10 percent of students moved between small and regular classes between grades. Krueger (1999) was able to gather data on actual random assignments from 1,581 students in 18 schools and found only a handful of instances (0.3 percent of students) in which a student was enrolled in a class size that differed from the one to which she was randomly assigned. To deal with class switching between grades, Krueger conducts an "intent to treat" analysis that defines treatment as the class type to which the student was initially assigned rather than the type actually attended. This analysis shows that class switching did not significantly bias the results.

Attrition from the data was also a significant issue in Project STAR. A large share of students who were randomly assigned to a class in the experiment do not appear in the data in

[12] Author's calculations from the NCES Common Core of Data state files.

later years because they left the school, repeated a grade, or skipped a grade. For example, among the 6,325 students that entered the experiment in kindergarten, 91 percent are in the kindergarten test data, 68 percent are in the first-grade data, 55 percent are in the second-grade data, and 47 percent are in the third-grade data. Krueger (1999) addresses this issue by crudely imputing missing test scores using the student's percentile score from the last year that they are observed in the data. This method produces qualitatively similar results to the analyses that exclude observations with missing test-score data. An analysis that uses more sophisticated methods for dealing with missing data, such as multiple imputation, could in theory yield somewhat different results. But this issue has largely been made moot by two recent follow-up studies that use administrative records and thus do not have significant attrition problems.

These two studies follow STAR participants into college and adulthood by matching students from the original experimental data to administrative records.[13] The first utilized IRS tax records to investigate a range of outcomes and found that students assigned to small classes at the beginning of elementary school are about two percentage points more likely to be enrolled in college at age 20, an impact that is statistically significant at the 10 percent level (Chetty et al., 2011). This study did not find any evidence of a class-size impact on students' incomes at age 27, but the income effects are measured with too much imprecision to warrant strong conclusions. Another contribution of this study was to verify that student assignment to class type was not correlated with a rich set of demographic information available in the IRS tax data, as the original dataset only included a small number of pre-treatment variables.

The second follow-up study utilized detailed college enrollment and completion data from the National Student Clearinghouse and found that students assigned to a small class were

[13] An earlier follow-up study found that test-score impacts (measured in percentiles) decreased after the class-size experiment ended and students returned to regular-size classes in grades 4–8, but that students who attended a small class in Project STAR were more likely to take a college entrance exam (Krueger & Whitmore, 2001).

about three percentage points more likely to attend college (Dynarski, Hyman, & Schanzenbach, 2011). The effect is largest among black students (5.8 percentage points) and students eligible for free lunch (4.4 points), but is not statistically significant from zero among white students and those students not eligible for the free lunch program. In other words, the substantial test-score gains found for blacks translated into higher college attendance rates but the somewhat smaller (but still significant) test-score impacts among whites did not. Dynarski, Hyman, and Schanzenbach (2011) also report that assignment to a small class in the early grades increased college degree attainment rates by about two percentage points.

In summary, researchers working with the STAR data have found positive effects of an early and very large reduction in class size on academic achievement in school and educational attainment. These are important results from a very strong research design. As noted previously, no other study in recent decades has randomly assigned students to smaller and larger classes in a substantial number of schools. The singularity of the STAR experiment is its Achilles heel in that, absent any other evidence from a randomized experiment, advocates and policymakers try to extrapolate more from the STAR results than is appropriate for any single experiment.

Like all experiments executed with at least a reasonable degree of fidelity, Project STAR produced credible estimates of the impact of a very specific intervention: being assigned to a class that was, on average, 8 students smaller in the early grades and then (usually) remaining in a class of that size through the end of the experiment (for students that did not switch schools). Because the small-class effect was concentrated in the first year students were in a small class, which was also the first year they attended that school, Hanushek (1999) raises the question of whether the class-size effect is mainly a socialization effect and not a more general benefit of smaller classes. Alternatively, it could be the case that the small-class effect in the first year would have dissipated over time in the absence of continued exposure to small classes. Absent a separate experiment in which students are randomly assigned to different size classes every year, it is impossible to resolve this question.

Hanushek (1999) also points out that large schools were overrepresented in Project STAR due to the decision to only include schools that had at least three classes per grade (to permit random assignment to all three class types), and that urban and predominantly minority schools were also overrepresented. For example, 37 percent of students in the STAR experiment were black, as compared to 21 percent of Tennessee children age 10–14 in the 1990 Census.[14] As a result, the overall class-size effect may be larger than would have been obtained with a representative sample of Tennessee elementary schools. Using the population shares by race to weight the effect estimates for blacks and whites rather than the STAR shares by race reduces the overall intent-to-treat effect by eight percent for test scores and 26 percent for college enrollment.[15]

A final issue regarding the STAR results that generalizes to almost any (hypothetical) class-size experiment is that teachers knew they were part of a study, the results of which might affect whether they would teach smaller classes in the future. Hoxby (2000) suggests that this incentive embedded in the STAR experiment may explain why such a large small-class effect was found. Krueger (1999) reports evidence that, within the regular-size classes in the STAR data, there is still a statistically significant association between class size and test scores. However, this result is difficult to interpret given that this source of variation in class size was

[14] Historical Population Data, Tennessee State Data, available at http://bus.utk.edu/cber/census/histcensus.htm.

[15] Specifically, the four-year effect on test scores, using results from Krueger (1999), is 0.18 when effects by race are averaged using the STAR weights and 0.17 when averaged using the 1990 Census weights. The impact on college enrollment, using results from Dynarski, Hyman, and Schanzenbach (2011), is 2.8 percentage points using the STAR weights and 2.1 points using the Census weights.

not randomly generated by the experiment but rather by other factors that may or may not be exogenous to student achievement.

QUASI-EXPERIMENTS BASED ON NATURALLY OCCURRING VARIATION IN CLASS SIZE

Quasi-experimental studies attempt to mimic randomized experiments by identifying variation in class size that is plausibly exogenous to student outcomes. The most credible of these quasi-experimental studies is Hoxby's (2000) examination of class-size variation in Connecticut that resulted from natural population variation triggering changes in the number of classes in a grade in a school. This study employs two distinct methods that produce similar results. The first method exploits changes in class size that results from idiosyncratic population changes. For example, a small school that has 15 first-grade students in one year and 18 the next year would have a larger class during the second year. The second method takes advantage of jumps in class size when a maximum class size rule is triggered. For example, a school that has set a class-size limit of 25 would have one second-grade class of 25 if there were 25 second-grade students but two classes of 13 if there were 26 students.

Hoxby (2000) finds no relationship between class size and achievement in fourth and sixth grade, which should reflect class size in all previous grades because the identification strategy uses variation that tends to be persistent within cohorts of students over time (and does not control for prior achievement). Hoxby's effects are what she calls "precisely estimated zeros"—in other words, even modest effects can be ruled out. Additionally, the Connecticut data do not provide any evidence of class-size effects at schools that serve disproportionately large shares of disadvantaged or minority students.

Hoxby argues that a significant advantage of her methodology is that teachers did not know that they were part of a study the results of which might influence their future working conditions. As discussed previously, it is difficult to assess this theory empirically using the experimental STAR data. Another important distinction between an explicit experiment like Project STAR and a study based on naturally occurring variation is that the experiment examines a well-defined treatment (e.g., “small” vs. “regular” classes) whereas natural variation occurs over a range of class sizes. Most elementary classes in the Connecticut study contained between 10 and 30 students (the mean class size was 21, with a standard deviation of 5.5 students). However, Hoxby (2000) does not find any evidence of class-size effects at any point in this range.

The only noteworthy limitation of the Connecticut data is that achievement tests are administered in the fall, so the set of students that make up the class size variable from a given school year will not be identical to the students who take the test in the fall of the following school year. Jepsen and Rivkin (2009) argue that this source of measurement error biases Hoxby’s results toward zero. However, significant attenuation bias seems unlikely given that within-school turnover is low in Connecticut—Hoxby reports that the average elementary school in 1997-98 had 93 percent of its students return.

Given the methodological strengths of the complementary methods employed by Hoxby (2000) in a study conducted more than a decade ago, it would seem that the same methods would have been applied using data from other states. Unfortunately this has not been the case, with only one exception. The primary challenge to researchers is obtaining data on class size by school and grade over a reasonably long time period. Most states collect data on school (and even school-by-grade) enrollment and test scores, but not on class size. Cho, Glewwe, and

Whitler (2012) address this limitation by obtaining historical class-size data from Minnesota through a survey administered to individual districts. They apply Hoxby's first method, using smooth changes in enrollment over time (not jumps due to maximum class size rules) to form instruments for class size.

An important limitation of the Minnesota study is that class-size data were obtained from only 22 percent of all districts for all years covered by the study, and from an additional 27 percent for some but not all years (a total of 52 percent). Cho, Glewwe, and Whitler present data indicating that the districts included in the analysis were similar in terms of observable characteristics to districts that were excluded due to missing data on class size, as well as evidence that measurement error in district reports of class size is unlikely to significantly bias their results. Missing data on a significant number of districts is less of a limitation in their study, which uses variation within schools over time, than in a study that uses across-district variation. But their results should still be interpreted with some caution given that they are based on data from only half of Minnesota districts.

Unlike Hoxby, Cho, Glewwe, and Whitler (2012) find positive effects of smaller classes, albeit of a smaller magnitude than the Project STAR effects. Specifically, their estimates imply that a reduction of class size by 10 students increases test scores in grades three and five by 0.04–0.05 standard deviations. (In the Minnesota study and most of the studies discussed below, it is important to bear in mind that reported effect sizes are based on a linear model of the relationship between class size and student outcomes—not an evaluation of an actual 10-student change in class size.) The estimated effect does not differ by race/ethnicity, gender, or free lunch eligibility.

The great advantage of the studies discussed so far is that they identify class-size effects using a source of variation in class size that is well understood. In the STAR experiment, assignment to a small class was done by lottery. In the Connecticut and Minnesota studies, effects were estimated using variation in class size that resulted from population variation. Many other studies simply examine naturally occurring variation in class size without focusing on a specific source of variation. The credibility of such studies is difficult to establish, so I do not review them here. The one exception is Rivkin, Hanushek, and Kain's (2005) study using longitudinal data from more than one-half million students in over three thousand schools in Texas during the 1990s.

Rivkin, Hanushek, and Kain (2005) control for student fixed effects as well as school-by-year fixed effects. Consequently, their class-size effects are estimated based off of differences in class size across different grades during the same year. This variation is not as plausibly exogenous as that resulting from population variation. But the authors argue that the variation comes from two sources: differences between cohorts of students in the number of transfers into or out of the school over time, and changes in school or district class-size policies.

Rivkin, Hanushek, and Kain (2005) find positive effects of smaller class sizes on reading and mathematics in fourth grade, a smaller but still statistically significant effect in fifth grade, and little or no effects in later grades. Because the researchers used a value-added model and state assessment results for which gain scores could only be computed beginning at fourth grade, they could not estimate class-size effects for the early grades that were studied in the STAR experiment. The estimated class-size effects for fourth- and fifth-grade students in Texas were generally in the range of 0.08–0.11 standard deviations per 10-student reduction in class size, with the exception of fifth-grade reading where the effect was only 0.03 standard deviations.

The results for sixth and seventh grade are all small and statistically insignificant, with the exception of an effect of 0.04 standard deviations (per 10-student reduction) in sixth-grade math. The estimated effects do not vary consistently by students' eligibility for free or reduced-price lunch.

All of the studies discussed so far focus on class size in elementary school, particularly in the early grades, with the exception of Rivkin, Hanushek, and Kain's (2005) inclusion of seventh-grade test scores. The only other credible study of class size in U.S. middle schools is Dee and West's (2011) analysis of eighth-grade students in the nationally representative National Education Longitudinal Study of 1988. Dee and West take advantage of the fact that students are observed in two subjects by comparing the outcomes of the same students who attended different size classes in different subjects. For example, they measure whether a student scores higher on a standardized mathematics test, on average, than on an English test if the math class was larger than the English class.

Dee and West (2011) find no overall impact of class size on test scores, i.e. the same students did not perform better in the subjects in which they had smaller classes. There was, however, a positive effect on test scores in urban schools, with a 10-student decrease in class size associated with an increase in test scores of 0.12 standard deviations (although the standard error implies a 95 percent confidence interval of approximately 0.03 to 0.21). The estimate for black students was similar in magnitude, but estimated with less precision and consequently statistically insignificant from zero.

Dee and West also found modest overall positive effects on non-cognitive skills related to school engagement. Students in smaller classes were less likely to say that they don't look forward to the subject, don't see it as useful, or are afraid to ask questions. Teachers of smaller classes said their students were less likely to be inattentive (but not more or less likely to be disruptive). Dee and West's findings are robust to conditioning on teacher fixed effects (i.e. controlling for the possible correlation between teacher quality and class size). They show that their measures of school engagement, like test scores, are correlated with long-term outcomes such as educational attainment and adult earnings, but their results are difficult to compare to the majority of studies (on both class size and other educational interventions) that focus on test scores and therefore ignore any effects through non-cognitive channels.

Most studies of class size are based on data from the U.S., and these studies are certainly of greatest interest to American policymakers. But given the relative paucity of evidence from the U.S., it is worthwhile to briefly review two international studies that provide credible evidence regarding the effects of class size. Angrist and Lavy (1999) took advantage of a class size limit in Israel of 40 students, just as Hoxby used various district-level class-size limits in Connecticut.[16] The Israel study finds positive effects of smaller fourth- and fifth-grade classes, with effect sizes indicating that a 10-student reduction in class size would raise student test scores by roughly 0.22 standard deviations in fifth-grade reading, 0.15 in fifth-grade math, and 0.10 in fourth-grade reading.[17] They do not find any effects on fourth-grade math scores or on third-grade scores in either subject. The general pattern of results is robust across a variety of specifications, but the magnitudes of the estimates vary and the analyses are each based on a single year of data so the results are not estimated with much precision. It is also important to

[16] Regression-discontinuity-based estimates of class-size effects may be biased if schools or families endogenously sort near the enrollment cutoffs. Urquiola and Verhoogen (2009) report evidence of such sorting using data from Chile, but it is not clear whether this issue applies in the context of the U.S. or other developed countries.

[17] These effect sizes are calculated by taking the coefficients in columns (2) and (8) of Tables IV and V in Angrist and Lavy (1999), multiplying by 10, dividing by the standard deviation of class mean scores, then multiplying by the ratio of between-class to total variation estimated using the third-grade micro-data (0.62).

note that the 40-student rule in Israel produced classes that tended to be far larger than those typical in the U.S.

Woessmann and West (2006), taking advantage of differences in average class size between the seventh and eighth grades within schools, examined class-size effects on performance on international examinations in 11 countries around the world. They find educationally meaningful effects of smaller classes in two countries, but no effects in most other countries. They are able to rule out large class-size effects in eight countries, and small effects in four countries.

Woessmann and West point out that the countries in which they find educationally meaningful positive effects of smaller classes are those with low salary levels for teachers (both on an absolute scale and relative to each country's per-capita GDP) and lower than average performance on international exams. A low average salary level for teachers suggests that a country is drawing its teaching population from a relatively low level of the overall capability distribution of all employees in this country. Thus the countries studied seem to have taken different paths, with some opting for relatively large numbers of poorly-paid teachers who perform better in smaller classes and others having relatively fewer but better-paid teachers whose performance isn't as affected by the number of students in class. However, Woessmann and West are limited in their ability to test this theory by the relatively small number of countries in their study (not to mention the challenges of inferring causality in any model where variation occurs at the country level).

STATEWIDE CLASS-SIZE POLICIES

In 2005, the Education Commission of the States identified 24 states that had adopted class-size reduction measures. The vast majority of states focused on class size in the early grades, usually defined as K–3 (but sometimes K–2 or K–4). Slightly more than half of states mandated smaller classes, usually by specifying a maximum class size for certain grades and subjects. For example, in 1986 the Louisiana legislature passed a law capping K–3 class size at 20. And in 1997, Alabama's state board of education set a timetable for schools to have no more than 18 students per teacher in grades K–3.

Other states have enacted voluntary policies that create incentives for schools to reduce class size—by providing state funding to do so—but do not require smaller classes. For example, in 1977 South Carolina provided additional funds to districts that attained an average pupil-teacher ratio of 21 in grades 1–3. The distinction between a mandate and an incentive is often blurry. For example, the Louisiana policy described above was to be enforced by denying funding for students above the 20-student limit, and specified that the measure could not take effect unless it was funded by the legislature.

States that provide districts with resources targeted at class-size reduction rather than additional unrestricted funding implicitly assume that districts will under-invest in class-size reduction if left to their own devices. This assumption is questionable given the broad popularity of smaller classes, and raises the question of whether forcing districts to reduce class size causes them to substitute away from more productive uses of the funding. There may well be political reasons for attaching the class-size label to state funding of education, or institutional features such as collective bargaining agreements that constrain districts' ability to pursue their preferred policies. But the implicit logic of statewide policies suggests an important reason why it may be inappropriate to use the results of studies of class-size experiments like Project STAR to support statewide class-size policies.

Only two statewide class-size initiatives, those in California and Florida, have been subjected to quasi-experimental impact evaluations. Both of those states enacted statewide class-size policies that make evaluation difficult by virtue of affecting all students, and consequently neither study is as credible as a randomized experiment or a strong quasi-experiment such as Hoxby (2000). But both studies make careful attempts to exploit variation in the extent of class-size reduction over time, and thus represent the best available evidence on the effects of large-scale class-size policies.

California's Class-Size Policy and its Unintended Consequences

In 1996, California enacted a K–3 class-size reduction program designed to reduce class size by ten students per class, from 30 to 20, throughout the state. Beginning in 1996-97, California provided $650 per pupil to schools that met class-size targets in grades K–3 (Sims, 2008). California did not administer any statewide exams until 1997–98, so evaluating this statewide policy is further complicated by the absence of any pre-program data coupled with the rapid adoption of smaller class sizes by many districts. Jepsen and Rivkin (2009) estimate the effect of the California policy by using variation at the school-grade-year level.[18] For example, class-size effects are identified off of changes in average class size in the same school-grade over time.

This method focuses on a relatively small proportion of schools because most schools fully implemented the policy in the first or second year and testing did not begin until the second

[18] Bohrnstedt and Stecher (2002) also evaluate the California policy and find inconclusive evidence regarding the policy's effect on student achievement. The authors argue that it is unclear whether the lack of evidence of positive policy effects is due to a truly small effect or limitations in the research design.

year. Jepsen and Rivkin (2009) report that 85 percent of schools participated in the class-size reduction program throughout the period of observation for grade 2, and 50 percent always participated in grade 3. As a result, the estimated effects are based only on data from the smaller share of late-adopting schools.

Holding teacher characteristics constant, Jepsen and Rivkin report that the 10-student reduction in class size increased test scores by 0.06–0.10 *school-level* standard deviations in math and 0.04–0.06 school-level standard deviations in reading. If one school-level standard deviation roughly corresponds to 0.5 student-level standard deviations, these effects correspond to effect sizes in student-level standard deviations of 0.03–0.05 and 0.02–0.03 in math and reading, respectively.[19] They find no evidence of effect heterogeneity by the racial composition of the school.

Effect estimates that hold teacher characteristics constant net out any effects of reduced teacher quality (to the extent that quality is correlated with observable characteristics). The California policy created a large number of new teaching positions, many of which were filled by new or not fully certified teachers. Jepsen and Rivkin (2009) find that having a first-year teacher (as opposed to a teacher with two or more years of experience) reduces student achievement by roughly the same amount as the 10-student reduction in class size increases achievement. In other words, students who ended up in the classrooms of teachers new to their classrooms and grades suffered academically from the teacher's inexperience by almost the same amount as they benefited from being in a smaller class. This result implies that the overall effect of the policy was smallest in the initial years when the number of new teachers was largest, and may have grown as the large crop of new teachers gained experience in the classroom.

[19] The conversion of 0.5 student-level standard deviations per school-level standard deviation is used by Sims (2009) and based on data from a large, diverse California school district.

Another concern surrounding large-scale class-size policies is that they will benefit advantaged students at the expense of disadvantaged students by creating teaching positions in affluent schools that will be filled by experienced teachers from disadvantaged schools. Jepsen and Rivkin (2009) find some evidence that the California policy initially helped advantaged students more than their less affluent peers, but that this difference dissipated in the long run as newly hired teachers gained in experience.

Reductions in teacher quality are the most cited unintended consequence of large-scale class-size policies, but Sims (2008, 2009) identified two others of consequence in California. The California policy created incentives for schools to create multi-grade classes. For example, a school with 30 first-grade students and 30 second-grade students could hire one additional teacher for a combination class that includes 10 students from each grade rather than hiring two additional teachers. Sims (2008) uses discontinuities in this incentive around the 20-student class-size maximum to instrument for the use of combination classes and finds large negative effects of being in a combination class, which more than offset any direct benefits of class-size reduction to students who were placed in combination classes as a result of the policy. These results imply a small overall effect of class-size reduction if the direct benefit of smaller classes is in the range found by Jepsen and Rivkin (2009). Sims (2008) also finds evidence that schools serving more disadvantaged students were more severely impacted by combination classes even though they were not significantly more likely to use them.

In a separate study, Sims (2009) finds that reducing first- and second-grade classes by about 10 students caused an average increase of two students per class in fourth and fifth grades—grades that were not covered by the policy. The unintended consequences of class-size reduction in California offer an important lesson: major education initiatives do not operate in a vacuum. Policies designed to affect one dimension of a student's educational experience are likely to affect others as well. Some of these might be thought of as implementation issues. For example, using combination classes to game the policy was not in keeping with the spirit of the policy.

But the rush to reduce class size in California, and its adverse impact on teacher quality, is a policy design issue that could easily have been avoided had the state legislature gradually increased incentives for class-size reduction rather than immediately encouraged a large reduction in class size. Jepsen and Rivkin (2009) simulate the long-term effects of the California policy, once any short-term impacts of teacher inexperience have dissipated. Converting their estimates to student-level standard deviations, they find long-term effects of the 10-student reduction in class size of 0.08 and 0.05 in math and reading, respectively.

Florida's K–12 Statewide Mandate

Six years after California enacted its policy that led to rapid reductions in class size in the early grades, Florida adopted an ambitious class-size reduction policy that covered all grades, at a cost of about $22 billion over the nine school years through 2011-12.[20] The Florida policy was implemented more gradually, perhaps in response to evidence of unintended consequences of California's policy.

Florida's policy was initiated in 2002, when voters narrowly approved (by a margin of 52 to 48 percent) an amendment to the state constitution that set limits on the number of students in core classes (such as math, English, and science) in the state's public schools. The Florida Constitution now includes a passage that requires public schools to assign no more than 18

[20] Florida Department of Education, Class Size Reduction Amendment web site, http://www.fldoe.org/classsize/.

students to each teacher in grades prekindergarten–3 by the beginning of the 2010–11 school year. The maximum class size is 22 students for grades 4–8 and 25 students for grades 9–12. In 2003, the Florida Legislature enacted a law that implemented the amendment by first requiring, from 2003–04 to 2005–06, districts to reduce their average class sizes either to the maximum for each grade grouping or by at least two students per year until they reached the maximum. Beginning in 2006–07, compliance was measured at the school level, with schools facing the same rules for their average class size that districts faced previously. Beginning in 2010–11, compliance was measured at the classroom level.

Taking advantage of the staggered introduction of class size reductions over time at the district and school level, Chingos (2012) utilized a comparative interrupted time series (CITS) research design to examine the effects of the policy on student achievement between 2004 and 2009. I used statewide student-level data to compare students who were more affected by the policy because they attended districts or schools that had pre-policy class sizes further from the mandated maxima to students that were less affected because they attended districts or schools that were already in compliance with the class-size policy. Specifically, I compared the deviations from prior trends in student achievement at districts/schools that were required to reduce class size to deviations from prior achievement trends at districts/schools that were not required to reduce class size.

In the analysis of the district-level implementation of the policy (2004–2006), I found that districts that were required to reduce class size in fact did so, over the course of three years, by 2–3 students more in grades 6–8 than districts not subject to this immediate requirement. The effect on middle school reading scores was negative and statistically insignificant, but precisely estimated enough to rule out small positive effects. The math results were closer to zero and not precise enough to rule out small positive effects.

The district-level results are not particularly surprising given that the relative reduction in class size was small, all districts received the same additional funding regardless of how much they had to reduce class size, and this analysis could only be applied to the middle grades, where previous research suggests class size effects are likely to be small or nonexistent.[21] My analysis of the first three years of the school-level implementation (2007–2009) of the policy is where one would expect to find evidence of class-size effects because the relative reduction in class size between treated and comparison schools was larger (3–4 students after three years), schools did not receive equal funding regardless of whether they needed to reduce class size, and the greater number of schools increases statistical precision.

But the school-level analysis does not yield any evidence of positive effects of class-size reduction; in fact, many of the point estimates are negative and statistically significant. In other words, even small positive effects can be ruled out for grades 3–5, which reflect class-size reductions both in those grades and persistent effects from earlier grades. In both the school- and district-level analyses, I did not find any evidence of heterogeneous effects, but in the school-level analysis I did find some evidence that class-size reduction improved non-cognitive outcomes, including student absenteeism in elementary school and incidents of crime and violence in middle school.

[21] The district-level analysis did not reveal any relative reduction in class size in the elementary grades in treated vs. comparison districts and consequently could not be used to estimate the policy’s impact on test scores in those grades.

COMPARING CLASS-SIZE EFFECT ESTIMATES

The Tennessee STAR experiment is the only randomized class-size evaluation performed at a significant scale, so naturally its results are often the measuring stick against which all other results are compared. The difficulty with such comparisons is that the STAR class-size effect varies dramatically depending on the time frame used to calculate it.[22] For example, test-score effects for kindergarten and first-grade students at the end of their first year in the experiment indicate that the 8-student reduction increased their test scores by about 0.2 standard deviations, or 0.03 standard deviations per 1-student reduction. That is a substantial effect. But the four-year effect for students who entered the study in kindergarten was 0.13, and the three-year effect for students who entered in first grade was 0.22, effect sizes which correspond to 0.005 and 0.01 standard deviations per 1-student reduction per year, respectively.[23]

The one-year effects from Project STAR are an outlier compared to the results of high-quality studies of naturally occurring variation in class size. Table 1 summarizes the results from the studies discussed above, scaling the effect sizes to correspond to a 10-student reduction in class size. The one-year STAR effect estimated by Krueger (1999) corresponds to 0.15 standard deviations per 10-student reduction in class size. This one-year effect is substantially larger than the effects found in all the other U.S. studies, such as Hoxby's (2000) precisely estimated zeroes in Connecticut and Cho, Glewwe, and Whitler's (2012) effects of 0.04–0.05 per 10-student reduction in Minnesota. This particular comparison is likely skewed in favor of the efficacy of smaller classes, as the Connecticut and Minnesota studies do not estimate one-year effects—rather, the effects should also reflect any lasting gains of smaller classes from previous years

[22] These kinds of comparisons also hinge on the interpretation of the large first-year effect in the STAR experiment. As discussed above, the design of the experiment does not allow for the estimation of a true one-year effect in the early grades, so it is impossible to know what would have happened, for example, had all first-grade students been returned to regular size classes.

[23] Calculations by the author using STAR public-use data.

(because the identification strategy compares cohorts of students who tend to consistently experience smaller or larger classes).

Rivkin, Hanushek, and Kain (2005) use a value-added model that will control for any effects of smaller classes from prior years, so the resulting estimates are most appropriately compared to per-year estimates from the STAR data. Krueger's (1999) four-year effect of 0.22 corresponds to an effect of 0.07 per 10-student reduction per year (dividing evenly over the four years). Assuming a larger first-year effect (as was found in STAR) implies a per-year effect of 0.04 per 10-student reduction in years two through four. These estimates are similar to Rivkin, Hanushek, and Kain's value-added estimates in fourth and fifth grade.

The evidence on class-size in middle school is less mixed than studies of the earlier grades, with the two high-quality studies that include students in grades 6–8 pointing to small or nil overall effects on test scores, although one study finds evidence of effects on outcomes other than test scores (Dee & West, 2011 and Rivkin, Hanushek, & Kain, 2005). There is not a single high-quality study of class size in high school, a gaping hole in the literature given that 30 percent of students are enrolled in these grades (Snyder & Dillow, 2012).

Evaluations of two statewide class-size policies produce estimates that generally fall short of the expectations generated by the results of the STAR experiment, but once again the comparison is affected by how one interprets the STAR results. The smallest predicted effect size from Project STAR would be obtained by ignoring the large first-year effect and using only the additional-year effect of about 0.005 standard deviations per one-student reduction in class size. This estimate is on the upper end of Jepsen and Rivkin's (2009) effect sizes of 0.002–0.005 per one-student reduction in California, and Chingos (2012) does not find any evidence of positive effects (and the Florida results are precise enough to rule out effects this small in some but not all grade-subject combinations).

Neither the Florida nor the California study finds evidence of larger effects for disadvantaged students, as was the case in the STAR experiment. In this respect the STAR experiment is also an outlier, as three additional studies did not find larger class-size effects for disadvantaged students (Cho, Glewwe, & Whitler, 2012; Hoxby, 2000; and Rivkin, Hanushek, & Kain, 2005). In the case of the statewide policies, it could be the case that the direct effects of smaller classes were larger for disadvantaged students, but the negative offsetting effects were larger as well. Jepsen and Rivkin (2009) find some evidence that disadvantaged students were more likely to be exposed to a new teacher as a result of the class-size policy.

The statewide policies also may not reflect the potential impacts of optimally designed class-size policies. The California policy might have produced better results had it been implemented more gradually with provisions in place to prevent the use of combination classes. The Florida policy might have been more successful had it targeted class-size reductions at certain grades or student populations rather than spreading additional resources thinly across all grades and students. However, given the mixed evidence of heterogeneous effects in the existing class-size studies it is not clear which students should be targeted. And it is also unclear whether funds are better targeted at the earlier or later grades—there is more evidence of class-size effects in the early grades but that could simply be an artifact of the dearth of studies of the middle and later grades.

The existing evidence also offers little guidance on what size classes should be targeted by policy. Data from Connecticut, Minnesota, and Texas indicate that most classes in the U.S. enroll between roughly 15 and 30 students (see Table 1). Will a reduction from 20 to 15 students have a larger impact than a reduction from 30 to 25 or 25 to 20? The STAR results were generated by a reduction from an average class size of 23 students to an average of 15 students, so classes of roughly 15 are sometimes targeted by policy (the federal policy had a goal of 18 and the Florida policy set a maximum of 18, both in the early grades). But several of the studies based on natural variation in class size include classes in this range and do not find class-size effects of the magnitude suggested by Project STAR. Hoxby (2000) specifically estimates models that allow for non-linear class-size effects and finds no evidence of such effects in Connecticut.

BENEFITS AND COSTS OF CLASS-SIZE REDUCTION

The evidence on the efficacy of class-size is clearly mixed, with one high-quality study finding quite large effects, another finding no effects, and a handful finding effects in between. But how do the costs of educating students in smaller vs. larger classes compare to the benefits? Put another way, what size class-size effects are needed to justify the cost of hiring additional teachers and building more classrooms?

The Tennessee STAR experiment generates the largest estimate of the payoffs of a large decrease in class size, so it is a natural starting point for a cost-benefit comparison. If the STAR results fail a cost-benefit test, then so will all other estimates of class-size effects because they are smaller whereas the costs are roughly the same. In Krueger's (1999) cost-benefit analysis, the return to the investment in smaller class sizes in Tennessee over the four-year experiment, in terms of expected increases in students' lifetime earnings, was slightly bigger than the costs of implementing the program.[24] If one were to apply this analysis to the other class-size studies,

[24] Krueger's (1999) method will understate the benefits of smaller classes to the extent that class-size reduction has positive impacts beyond those on students' future earnings.

which find smaller benefits, it is unlikely that the benefits would exceed the costs. For example, Cho, Glewwe, and Whitler (2012) apply Krueger's cost-benefit analysis to the Minnesota results and find that the costs of an eight-student reduction in class size are almost six times the expected benefits (in terms of future earnings).

It is particularly difficult for class-size reduction to pass a simple cost-benefit test such as Krueger's because it is so expensive. Krueger (1999) estimates that reducing class size by one third increases per-pupil costs by one third. In the most recent OECD (2011) data, average class size in U.S. primary schools was 23.8 and per-pupil spending was $9,982. Krueger's calculation implies that reducing class size by one third, from about 24 to about 16, would cost $3,327 per student. This translates to $421 in per-pupil spending for each one-student reduction in class size.

This is likely an upper bound estimate of the cost of class-size reduction because it assumes that all costs are variable, which is true for the majority of the costs of education (e.g., teachers and classrooms) but not all (e.g., administrators and transportation, assuming no changes in the administrator and bus driver to student ratio). Whitehurst and Chingos (2011) estimate teacher salary costs of a one-student reduction in the pupil-teacher ratio of about $250 per pupil. Roza and Ouijdani (2012) estimate per-pupil cost savings from increasing class size of $161 per one-student reduction. Both of these estimates understate the cost implications of changing class size because they do not account for teacher benefits or space costs. Harris (2009) includes estimates of teacher benefits and capital costs and arrives at a per-pupil estimate of $223 per one-student reduction in class size.

It is obvious that a policy should not be pursued unless it has benefits that are greater than its costs, although of course it is difficult to accurately measure benefits. But when resources are limited, passing this test is necessary but not sufficient. The cost-benefit test any educational policy must pass is not “Does this policy have positive effects that justify its costs?” but rather “Is this policy the most productive use of these educational dollars?” Assuming even the largest class-size effects, such as the STAR results, class-size decisions must still be considered in the context of alternative uses of tax dollars for education. Will a dollar spent on class-size reduction generate as much return as a dollar spent on raising teacher salaries, implementing better curriculum, strengthening early childhood programs, providing more frequent assessment results to teachers to help guide instruction, or making investments in educational technology?

There is no research from the U.S. that directly compares class size to specific alternative investments. In other words, the comparison conditions for nearly all class-size studies has been smaller vs. larger classes rather than, for example, a comparison of $20 billion invested in smaller classes vs. $20 billion invested in higher teacher salaries. Thus estimates of effects and costs from different education investments have to be extrapolated and estimated from different studies, and this process is necessarily inexact. Nevertheless, Harris (2009) finds short-term rates of return for computer-aided instruction, cross-age tutoring, early childhood programs, and increases in instructional time that are all greater than those for class-size reduction. Harris’s analysis is important because it represents the kind of careful comparison of costs and benefits that is too rarely undertaken in education research.

Dynarski, Hyman, and Schanzenbach (2011) also compare the costs and benefits of various interventions by estimating the cost per student induced into college. For example, a program that costs $10,000 per student and increases college enrollment rates by 5 percentage points is estimated to cost $200,000 per student induced into college. Dynarski et al. estimate that class-size reduction in Project STAR cost about $400,000 per student induced into college.

This is on the upper end of the programs which are used for comparison, including Head Start ($133,000), Abecedarian ($410,000), the Social Security Student Benefit Program ($21,000), and helping low-income families apply for college financial aid ($1,257).

THE NEED FOR MORE RESEARCH

More than 70 years ago, Douglass and Parkhurst (1940) summarized the debate over class size as follows:

> Those who advocate larger classes in the schools maintain that such classes are preferable because they offer a better socializing and democratizing situation and contribute to a lighter load in terms of class periods per week. These advocates are not unmindful of possible reduced expenditures as a result of large classes. On the other hand, the many who strongly advocate small classes hold that the most important outcomes of teaching, such as character development and appreciations, cannot be acquired by pupils in large classes as readily as in small classes; and that these outcomes, not being measured, are overlooked in comparisons. They claim that the large class is a false economy.

The debate has changed very little since then. Research on class size is decidedly mixed and offers little guidance as to what grades, students, and range of class sizes represent opportunities for cost-effective investments. An advocate with a conclusion in search of a study can surely find one, even just among the set of high-quality studies reviewed here.

This state of affairs is unlikely to be satisfying to school administrators and state lawmakers who want to know what the optimal class-size policy is for their district or state. Of course research will never reveal the optimal set of educational policies for every context, but there is much more that could be learned about the effect of class size on student outcomes.

First, researchers should take advantage of the student-level education databases that most states have developed over the last decade to study class size. Methods similar to those employed by Rivkin, Hanushek, and Kain (2005) using data from Texas could easily be applied to similar data from other states. States that have had data systems in place for a long enough period of time could also be studied using the more credible quasi-experimental methods of Hoxby (2000), but with the ability to more accurately capture effect heterogeneity by examining subgroups of students rather than subgroups of schools. The primary data limitation that exists in many states is the absence of administrative data on class size, but such data could be collected from districts, as was done in Minnesota by Cho, Glewwe, and Whitler (2012).

Second, given the billions of dollars that state governments have invested in class-size reduction it is unconscionable that Project STAR remains the only randomized experiment of class size conducted at a significant scale. In the current fiscal environment it is unlikely that states are going to fund experimental class-size reduction efforts. But given that many states are being forced to make budget cuts, those cuts could be implemented in ways that are conducive to evaluation. For example, a state that provides financial support related to class size could phase in budget cuts to that program using random assignment at the district level or an index of student disadvantage that lends itself to a regression discontinuity research design.

Third, states could pilot programs that will enable research on the relative efficacy of class-size reduction compared to other policies that require similar resources. The class-size mandates that many states have adopted may constrain schools from maximizing the productivity of the taxpayer dollars invested in public education, but there is scant evidence on this question. Few studies have compared class-size reduction to other uses of the same resources. A state with a class-size mandate could solicit applications from districts that wish to exceed the mandated caps on class size and reinvest the resources in other programs, and then randomly assign when a district is granted such a waiver (e.g., the upcoming academic year or the following year).

Another type of program a state might pilot would replace across-the-board mandates or incentives with a policy that would target certain kinds of teachers for assignment to smaller classes, such as new teachers with weak classroom management skills. Such a policy could be enacted by a state, for example by giving superintendents or principals a pot of money to use for discretionary class-size reduction. But it might provoke protests from teachers who do not get a smaller class, and from the parents of their students.

Finally, research on class size should be one part of a larger discussion on how to maximize the productive use of school personnel. For example, the subject of class scheduling practices highlights that class sizes are not set in a vacuum. All else equal, a decrease in the number of class periods a teacher is assigned to teach will produce a decrease in the pupil-teacher ratio without changing class size because more teachers will be required for the same number of students. This applies to elementary schools that use some amount of pull-out instruction, but will be particularly important in high schools that deliver subject-specific instruction.

Figure 2 shows National Education Association data on both average class size in secondary schools (reproduced from Figure 1) alongside the number of students taught per day by the average teacher. As discussed previously, the average class size in these data bounced around somewhat between 1960 and 2005 but did not change markedly over this period. But the number of students taught per day fell dramatically from about 130 in the late 1960s and early 1970s to around 85 in the early 2000s—a decrease of approximately one third. In other words, high schools would need to have hired about 50 percent more teachers in order to educate the same number of students in the same size classes.

This trend indicates that teaching load—the number of students a teacher is responsible for—is a potentially important factor to consider alongside class size. For example, class size could be reduced without hiring more teachers by requiring existing teachers to teach for a larger number of periods in the day. That the recent historical pattern is the opposite—constant class size and decreased teaching load—suggests that the number of periods spent giving instruction has declined. Recent research has largely ignored the subject of teaching load, probably because most studies have focused on the elementary grades where teachers in self-contained classrooms face teaching loads and class sizes that are usually equal.

Teaching load and class size are part of a larger set of decisions that schools make about how to organize their instructional personnel. Are students in elementary schools best educated by a single classroom teacher or by multiple teachers with strengths teaching particular subjects? Is the traditional model of a single teacher in the classroom superior or inferior to team teaching, in which two teachers are jointly responsible for a class? Should high-school students be taught in 45-minute periods or in 90-minute blocks? How many class periods should be set aside for teachers to prepare lesson plans?

New evidence on this broader set of questions might identify ways to more efficiently allocate the time of school personnel so as to reduce expenditures without class size increasing in lock step.

CURRENT POLICY DEBATES

The popularity of smaller classes may make it politically difficult for policymakers to increase class size in order to sustain other investments in education, even in a time of budget austerity. However, there is some evidence that teachers and the public in general may be open to modest increases in class size in order to allow for other investments. In a 2006 survey, 83 percent of teachers in Washington state said they preferred a salary increase of $5,000 to a two-student reduction in class size (Goldhaber, DeArmond, & DeBurgomaster, 2010). Most recently, in a nationally representative survey of Americans, 73 percent of respondents said they preferred a class of 27 students taught by one of the best teachers in their school district to a class of 22 students taught by an average teacher (Farkas & Duffett, 2012).

These survey results provide suggestive evidence that some changes to existing class-size policies might be politically feasible. In this context, state policymakers might consider amending class-size policies to provide local school leaders more flexibility in how to distribute support for smaller classes. Much smaller classes for inexperienced teachers who need support in developing classroom management skills or for teachers who are responsible for struggling students may make more sense than across-the-board reductions. States might even allow districts to apply for waivers that would allow them to spend the funds on purposes other than class-size reduction that they believe are more cost-effective.

The tradeoff between class size and teacher salaries needs to be very carefully considered. Effects on student achievement related to differences in teacher quality are very large. A recent review of research on teacher value-added finds that having a teacher who is one standard deviation above average (as compared to the average teacher) increases test scores by 0.11 and 0.15 standard deviations in reading and math, respectively (Hanushek & Rivkin, 2010). These short-term impacts on test scores translate into better outcomes later in life, including higher earnings and an increased likelihood of attending college (Chetty, Friedman, & Rockoff, 2011).

With fixed or reduced state budgets to support K–12 education, maintaining class-size limits means a larger pool of teachers with lower salaries. It means that funds that might be devoted to raising teacher salaries across the board or selectively in hard-to-fill positions or for highly effective teachers will be limited. By one estimate, an increase in average class size by 5 students would result in an across the board increase of 34 percent in teacher salaries if all the savings were devoted to that purpose (Chingos, 2011). Higher salaries would likely draw more qualified people into the teaching profession, and keep them there.

In the current fiscal climate, it is clear that the yearly increases in funding in real dollars that have long been enjoyed by our nation's public schools are coming to an end for the foreseeable future. Many states and districts are contemplating cuts in funding that will require schools to make hard choices. So although the research literature has focused on the effect of reducing class size, the current policy debate concerns the other side of the coin—the consequences of increasing the size of classes. The potential for negative consequences of larger classes clearly needs to be weighed against the fallout from cutting other programs in order to preserve smaller classes, including both academic programs and non-academic offerings such as athletics and the arts.

An important related consideration is that the effect of any increase in class size will depend on how such an increase is implemented. One rough calculation discussed earlier indicated that a one-student increase in the pupil-teacher ratio in the U.S. would decrease the teaching workforce by about 7 percent (Whitehurst & Chingos, 2011). Many school districts and states across the nation are considering reductions in the teacher workforce on this order of magnitude. If the teachers to be laid off were chosen in a way largely unrelated to their effectiveness, such as "last in first out," then the associated increase in class size could well have a negative effect on student achievement. But if schools choose the least effective teachers to let go, then the effect of increased teacher quality could make up for some or all of any negative effect of increasing class size (see, e.g., Boyd et al., 2011 and Goldhaber & Theobald, 2010).

State resources for education should always be carefully allocated, but the need to carefully weigh costs and benefits is particularly salient in times of austere budgets. The conventional wisdom that class-size reduction "works," and does so especially well for disadvantaged students in the early grades, is based primarily on Project STAR, which is a single study (albeit a very important one) from a single state in the 1980s. The number of other high-quality class size studies is small, but they collectively indicate that the STAR results are an outlier and that the likely benefits of smaller classes—or the likely harm of larger classes—are substantially smaller, and in some contexts may be negligible. These other studies also do not find consistent evidence of particularly large impacts in certain grades or among specific types of students.

There is clearly a need for more research, but the weight of the existing high-quality evidence indicates that although smaller classes may represent a cost-effective investment in some circumstances, many school systems in the U.S. have overinvested in class-size reduction. In other words, there are likely many circumstances in which modest increases in class size would benefit students if the resources were reinvested in more cost-effective interventions, such as those identified by Harris (2009) and Dynarski, Hyman, and Schanzenbach (2011). In times of decreasing educational resources, the strong link between class size and spending coupled with the evidence on cost-effectiveness means that small increases in class size are likely to offer a path to balanced budgets that minimizes harm to students.

Acknowledgments

I thank Beth Akers, Doug Besharov, Mike Gallaher, Russ Whitehurst, and three anonymous reviewers for helpful comments. Any remaining errors are my own.

References

Angrist, J.D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics, 114, 533–575.

Bohrnstedt, G.W., & Stecher, B.M. (Eds.). (2002). What we have learned about class size reduction in California. Palo Alto, CA: CSR Research Consortium.

Boyd, D.J., Lankford, H., Loeb, S., & Wyckoff, J.H. (2011). Teacher layoffs: An empirical illustration of seniority versus measures of effectiveness. Education Finance and Policy, 6, 439–454.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. Quarterly Journal of Economics, 126, 1593–1660.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper No. 17699. Cambridge, MA: National Bureau of Economic Research.

Chingos, M.M. (2011). The false promise of class-size reduction. Washington, DC: Center for American Progress.

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

Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. Economics of Education Review, 31, 77–95.

Coopersmith, J. (2009). Characteristics of public, private, and Bureau of Indian Education elementary and secondary school teachers in the United States: Results from the 2007-08 Schools and Staffing Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Dee, T.S. (2004). Teachers, race, and student achievement in a randomized experiment. The Review of Economics and Statistics, 86, 195–210.

Dee, T.S., & West, M.R. (2011). The non-cognitive returns to class size. Education Evaluation and Policy Analysis, 33, 23–46.

Douglass, H.R., & Parkhurst, A.J. (1940). Size of class and teaching load. Review of Educational Research, 10, 216–221.

Dynarski, S., Hyman, J., & Schanzenbach, D.W. (2011). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. NBER Working Paper No. 17533. Cambridge, MA: National Bureau of Economic Research.

Education Commission of the States. (2005). State class-size reduction measures. Denver, Colorado: Education Commission of the States.

Farkas, S., & Duffett, A. (2012). How Americans would slim down public education. Washington, DC: Thomas B. Fordham Institute.

Finn, J.D., & Achilles, C.M. (1990). Answers and questions about class size: A statewide experiment. American Educational Research Journal, 28, 557–577.

Folger, J., & Breda, C. (1989). Evidence from Project STAR about class size and student achievement. Peabody Journal of Education, 67, 17–33.

Glass, G.V., & Smith, M.L. (1979). Meta-analysis of research on class size and achievement. Educational Evaluation and Policy Analysis, 1, 2–16.

Goldhaber, D., DeArmond, M., & DeBurgomaster, S. (2010). Teacher attitudes about compensation reform: Implications for reform implementation. CALDER Working Paper 50. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D. & Theobald, R. (2010). Assessing the determinants and implications of teacher layoffs. CALDER Working Paper 55. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Hanushek, E.A. (1999). The evidence on class size. In S.E. Mayer & P.E. Peterson (Eds.), Earning and learning: How schools matter. Washington, DC: Brookings Institution Press.

Hanushek, E.A. (2003). The failure of input-based schooling policies. Economic Journal, 113, F64–F98.

Hanushek, E.A., & Rivkin, S.G. (1997). Understanding the twentieth-century growth in U.S. school spending. Journal of Human Resources, 32, 35–68.

Hanushek, E.A., & Rivkin, S.G. (2010). Generalizations about using value-added measures of teacher quality. American Economic Review, 100, 267–271.

Harris, D.N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. Educational Evaluation and Policy Analysis, 31, 3–29.

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

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: the potential tradeoff between teacher quality and class size. Journal of Human Resources, 44, 223–250.

Keaton, P. (2012). Public elementary and secondary school student enrollment and staff counts from the Common Core of Data: School year 2010–11. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Krueger, A.B. (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, 115, 497–532.

Krueger, A.B. (2003). Economic considerations and class size. Economic Journal, 113, F34–F63.

Krueger, A.B., & Whitmore, D.M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. The Economic Journal, 111, 1–28.

Millsap, M.A., Giancola, J., Smith, W.C., Hunt, D., Humphrey, D.C., Wechsler, M.E., & Riehl, L.M. (2004). A descriptive evaluation of the federal class-size reduction program: Final report. Washington, DC: U.S. Department of Education.

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, 165–177.

National Education Association. (2010). Status of the American public school teacher 2005–2006. Washington, DC: National Education Association.

Nye, B., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? Educational Evaluation and Policy Analysis, 26, 237–257.

OECD. (2011). Education at a glance 2011: OECD indicators. Paris, France: OECD Publishing.

Oliff, P., Mai, C., & Palacios, V. (2012). States continue to feel recession’s impact. Washington, DC: Center on Budget and Policy Priorities.

Ritter, G.W., & Boruch, R.F. (1999). The political and institutional origins of a randomized controlled trial on elementary school class size: Tennessee’s Project STAR. Educational Evaluation and Policy Analysis, 21, 111–125.

Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. Econometrica, 73, 417–458.

Rockoff, J. (2009). Field experiments in class size from the early twentieth century. Journal of Economic Perspectives, 23, 211–230.

Roza, M., & Ouijdani, M. (2012). The opportunity cost of smaller classes: A state-by-state spending analysis. Unpublished paper.

Sims, D. (2008). A strategic response to class size reduction: combination classes and student achievement in California. Journal of Policy Analysis and Management, 27, 457–478.

Sims, D. (2009). Crowding Peter to educate Paul: Lessons from a class size reduction externality. Economics of Education Review, 28, 465–473.

Snyder, T.D. (Ed.). (1993). 120 years of American education: A statistical portrait. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Snyder, T.D., and Dillow, S.A. (2012). Digest of education statistics 2011. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. American Economic Review, 99, 179–215.

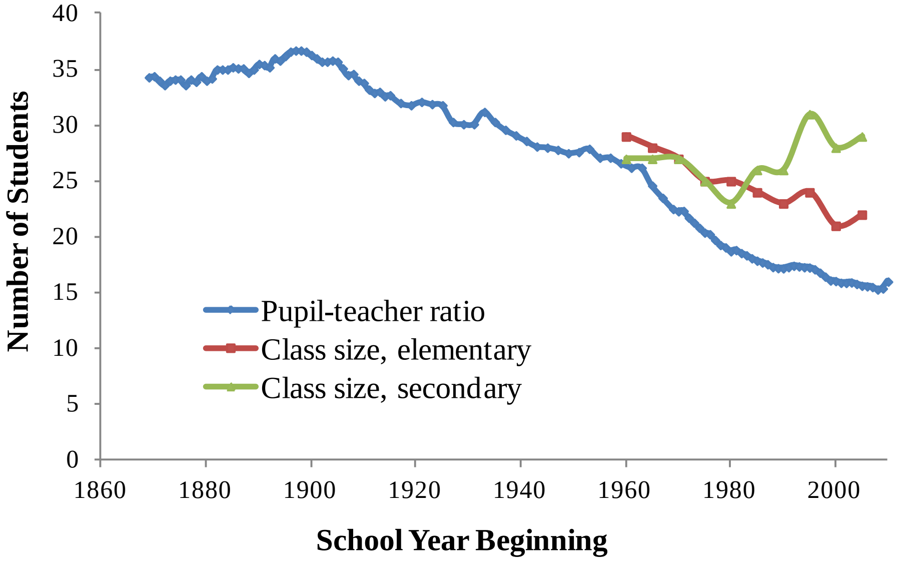
Whitehurst, G.J., & Chingos, M.M. (2011). Class size: What research says and what is means for state policy. Washington, DC: Brookings Institution.

West, M.R., & Woessmann, L. (2006). Which school systems sort weaker students into smaller classes? International evidence. European Journal of Political Economy, 22, 944–968.

Woessmann, L., & West, M. (2006). Class-size effects in school systems around the world: evidence from between-grade variation in TIMSS. European Economic Review, 50, 695–736.

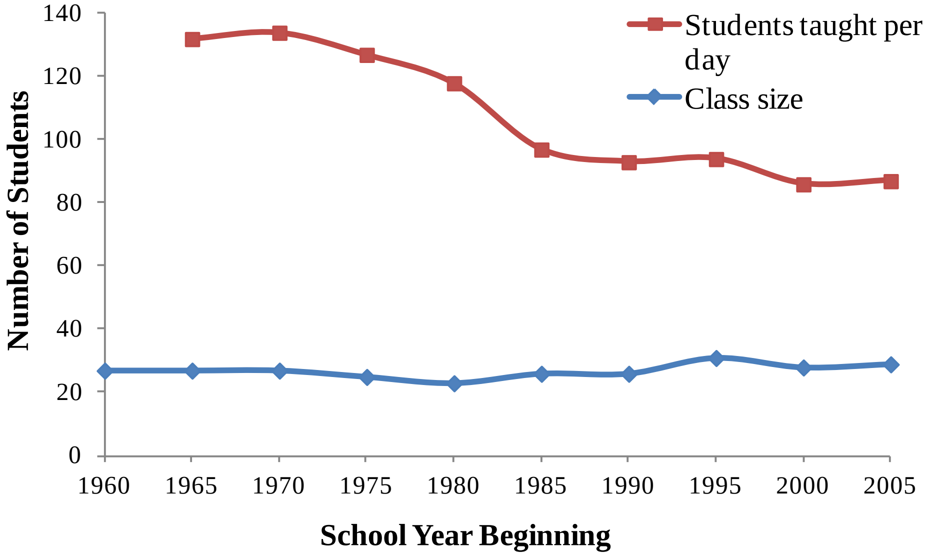
Word, E., Johnston, J., Bain, H.P., Fulton, B.D., Zaharias, J.B., Lintz, M.N., Achilles, C.M., Folder, J., & Breda, C. (1990). Student/teacher achievement ratio (STAR): Tennessee's K–3 class size study, final report. Nashville, TN: Tennessee State Department of Education.

Figure 1: Pupil-Teacher Ratio and Class Size in U.S. Public Schools, 1860–2010



Sources: Keaton (2012), Table 4; Snyder and Dillow (2012), Table 69; Snyder (1993), Table 14; National Education Association (2010), Tables 20 and 22.

Figure 2: Class Size and Students Taught per Day, Secondary Schools, 1960–2005



Source: National Education Association (2010), Tables 22 and 23.

Table 1. Summary of High-Quality Evidence on Class Size Impacts on Test Scores

Study	Data	Variation in class size	Effect on student test scores, in student-level standard deviations	Effect heterogeneity
Angrist and Lavy (1999)	Class-level, Grades 4–5, Israel, 1991–92	Mean of 30–31, standard deviation of 6–7	About 0.22 per 10-student reduction in grade 5 reading; 0.15 in grade 5 math; 0.10 in grade 4 reading; insignificant estimated in grade 4 math; no effect in grade 3.	Suggestive evidence of larger effects for disadvantaged students.
Chingos (2012)	Student-level, Grades 3–8, Florida, 2001–2009	Relative reduction of 2–4 students in quasi-experimental treatment group relative to comparison group (from mean of 23–25)	No effects; small effects can generally be ruled out	No consistent evidence of heterogeneity by student race/ethnicity, eligibility for free lunch, or gender.
Cho et al. (2012)	School-grade-level, Grades 3 and 5, Minnesota, 1997–2005	Mean of 22–24, standard deviation of 5	0.04–0.05 per 10-student reduction	No heterogeneity by school-level racial, gender, or free lunch composition.
Dee and West (2011)	Student-level, Grade 8, United States, 1988	Mean of 25, standard deviation of 6	0.02–0.03 per 10-student reduction (not statistically significant)	0.12 effect of 10-student reduction in urban schools (0.045 standard error)
Hoxby (2000)	School-grade- and district-grade-level, Grades 4 and 6, Connecticut, 1986–98	Mean of 21–24, standard deviation of 5–6	No effects; precisely estimated zeros	No heterogeneity by school-level average income or percent black.

Table 1 (continued). Summary of High-Quality Evidence on Class Size Impacts on Test Scores

Study	Data	Variation in class size	Effect on student test scores, in student-level standard deviations	Effect heterogeneity
Jepsen and Rivkin (2009)	School-grade-level, Grades 2–4, California, 1997–2001	Policy reduction from 30 to 20	Effects of 10-student reduction: 0.03–0.05 in math and 0.02–0.03 in reading	No heterogeneity by school racial composition.
Krueger (1999)	Student-level, Grades K–3, Tennessee, 1986–89	Treatment group with mean of 23, control group with mean of 15	8-student reduction: 0.12 after one year, 0.22 after four years. Effects scaled to correspond to 10-student reduction: 0.15 after one year, 0.28 after four years	Four-year effect (of 8-student reduction): 0.20 free lunch eligible, 0.15 not free lunch eligible, 0.24 black, 0.15 white, 0.31 inner city, 0.15 Metropolitan, 0.20 rural
Rivkin et al. (2005)	Student-level, Grades 4–7, Texas, 1993–95	Mean of 20–23, standard deviation of 2–4	0.08–0.11 per 10-student reduction in grade 4–5 math and grade 4 reading; 0.03-0.04 in grade 5 reading and grade 6 math; no effect in grade 6 reading or grade 7.	No heterogeneity by student eligibility for free or reduced-price lunch
Woessmann and West (2006)	Student-level, Grades 7–8, 11 Countries, 1994–95	Varies by country: means of 20–33, standard deviations of 3–13	Effects on pooled math and science scores: 0.20–0.22 per 10-student reduction in Greece and Iceland; effects of 0.30 ruled out in 8 countries; effect of 0.10 ruled out in 4 countries	