Journal of Economic Perspectives—Volume 23, Number 4—Fall 2009—Pages 211–230

Field Experiments in Class Size from the Early Twentieth Century

Jonah Rockoff

A vast majority of adults believe that class size reductions are a good way to improve the quality of public schools. For example, the Gallup Poll of the Public's Attitudes Toward the Public Schools (an annual survey representative of the non-institutionalized U.S. population age 18 and over) asked respondents in 1999 what they would do to improve public schools if they could change one thing; reducing class size was the most prevalent choice among parents of public school children and the second most prevalent choice (after increased discipline) among individuals without children.

Reviews of the research literature, on the other hand, have provided mixed messages on the degree to which class size matters for student achievement (for example, Glass and Smith, 1978; Hanushek, 1986; Hedges, Laine, and Greenwald, 1994; and Krueger, 2003). The literature since World War II has followed two main tracks. A wave of research on class size from the 1950s into the 1990s typically used modern statistical methods to analyze large datasets containing information on a variety of educational inputs and outcomes. However, in the last decade or two, economists have focused on the difficulty of using observational data on class size to draw inferences about causal relationships. After all, class size is likely to be correlated with a number of factors not observable to the researcher, like the willingness of parents and the local community to support education in many ways, and so the specific effect of class size may be hard to identify.

Thus, researchers turned to studying experimental and quasiexperimental variation in class size (for example, Angrist and Lavy, 1999; Krueger, 1999; Hoxby,

■ *Jonah Rockoff is Associate Professor of Economics and Finance, Graduate School of Business, Columbia University, New York City, New York, and Faculty Research Fellow, National Bureau of Economic Research, Cambridge, Massachusetts. His e-mail address is* ⟨*jonah.rockoff@columbia.edu*⟩.

2000). Tennessee's Student Teacher Achievement Ratio Project—"Project STAR," as it is known—is the most prominent example. The largest class size experiment ever conducted, it involved the participation of 79 schools, 328 classes, and roughly 6,500 students when it began in 1985. Students and teachers were randomly assigned to small or large classes (averaging 15 or 23 students, respectively) from kindergarten through third grade, and students have been followed over time. The numerous studies sparked by Project STAR present persuasive evidence that major class size reductions at early ages can have significant short- and long-term benefits (for example, Bain and Achilles, 1986; Word et al., 1990; Krueger, 1999; and Krueger and Whitmore, 2001). It is difficult to overstate the impact of this research on the class size debate and support for class size reduction policies.

Far less well known is a substantial body of experimental work on class size that developed prior to World War II. These field experiments did not have the benefit of modern econometrics and only a few were done on a reasonably large scale. However, they often used careful empirical designs, and the collective magnitude of this body of work is considerable. Moreover, this research produced little evidence to suggest that students learn more in smaller classes, which stands in contrast to some, though not all, of the most recent work by economists.

In this essay, I provide an overview of the scope and breadth of the field experiments in class size conducted prior to World War II, the motivations behind them, and how their experimental designs were crafted to deal with perceived sources of bias. I conclude with a discussion of how one might interpret the findings of these early experimental results alongside more recent research, and how research on class size has shifted towards using instrumental variables rather than field experiments to address the class size issue empirically.

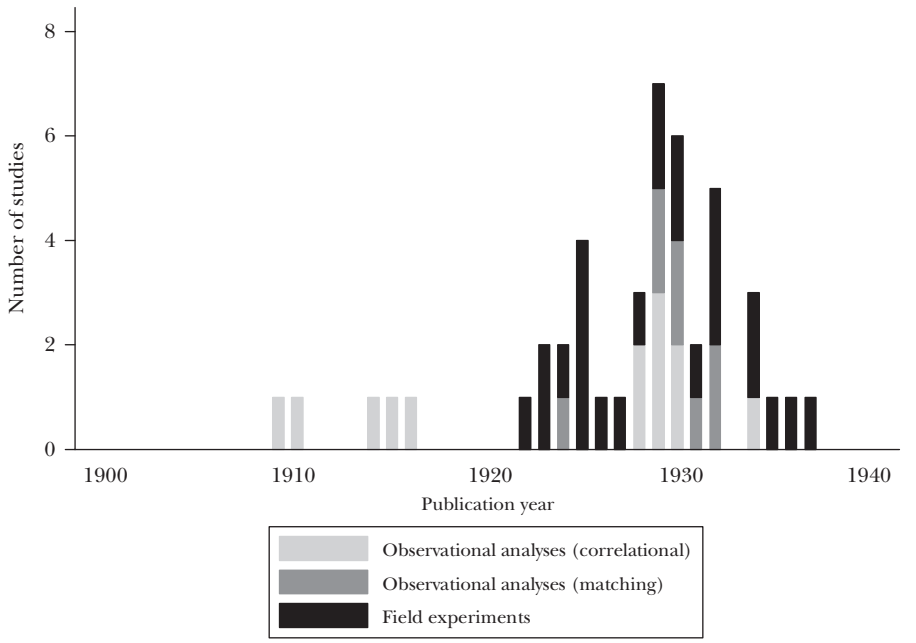
A Wave of Class Size Research

> *Larger classes make for fewer teachers and lower building costs. Increasing the size of classes, then, offers an obvious and tempting means to immediate educational economy.*
>
> Earl Hudelson (1928), University of Minnesota

Forty-five studies of the impact of class size on student achievement were published prior to 1940 (see Figure 1).[1] While only a handful of such studies surfaced before 1920, publications surged over the next decade, peaked in 1929, and trailed off in the mid-1930s. Thirteen of these studies (including all of those published before 1920) were correlational analyses of observational data, and eight used observational data but employed matching methods to control for student characteristics (that is, each student in a smaller class was paired with a similar

[1] These 45 studies include only the analyses of primary data that focused on the effect of class size, and exclude reports of summary statistics, literature reviews, opinion pieces, and criticism of other research. Full citations are provided in an online appendix to this paper available at ⟨http://www.e-jep.org⟩.

Figure 1
Studies of the Impact of Class Size Published Prior to 1940



student in a larger class, and outcomes for the paired students were analyzed). The remaining 24 studies were field experiments, with research designs that involved active assignment of pupils and teachers to classrooms of different sizes.

To give a better sense of the scope of this work, I have taken the 24 field experiments and broken them down along a number of dimensions, presented in Table 1. Studies were conducted in 13 states and involved the participation of roughly 80 schools, 650 classrooms, and 17,000 students in a wide range of educational settings. Seventeen studies examined class size at the high school level, including roughly 13,000 students. Fewer but still significant numbers of investigations were made in grades K–3 (two studies, 1,000 students), grades 4–8 (five studies, 1,800 students), and at the university level (two studies, 1,200 students).[2] By far, the most common subject areas examined were English (13 studies) and mathematics (ten studies). However, researchers examined performance in eight other areas, including science, social studies, history, and Latin. Test scores were used to measure student achievement in all studies except Davis (1923), where letter grades based on high school students' final exams were used.

[2] One study (Metzner and Berry, 1926) focused on mentally handicapped children for whom grade level was not given. Studies that included pupils from two or more grade categories are counted once in each category, so the total sums to more than 24. Studies are also counted in this manner for statistics on subject area.

Table 1
Summary Statistics of Field Experiments in Class Size through 1940

Author (Year of Publication)	*State(s)*	*Grades covered*	*Subjects covered*	*Schools*	*Classes*	*Students*	*Average class size Small/Large*
Stevenson (1922,1923)	IL	2	Math, English+	5	6	288	37/48
		5	Math, English+	5	6	298	40/50
		7	Math, English+	5	4	148	39/47
		9, 10	Math, English, Science, History, Latin, French	4	134	3845	21/37
Almack (1922)	CA	5, 7	Math, English	1	8	96	12/28.3
Davis (1923)	20 Cities in the Midwest	H.S.	Math, English	20	240	6130	20/30
Tope, Groom, & Beeson (1924)	CO	11	English	1	3	98	20/39
Averill & Mueller (1925)	MA	5	Reading	1	2	38	12/26
Stevenson (1925a)	OH	2	Math, English+	25	26	657	25/42
		5	Math, English+	25	20	525	26/44
		7	Math, English+	25	16	404	25/41
Stevenson (1925b)	MA, KY, MI	H.S.	Math	3	10	275	18/37
Metzner & Berry (1926)	MI	Elementary	Special Education	*n/s*	24	534	15/30
Hudelson (1927)	MN	Higher Ed	Multiple subjects	1	14	693	19/80
Haertter (1928)	MN	H.S.	Plane Geometry	1	2	75	20/55
Judd (1929)	*n/s*	H.S.	Geometry, Spanish	2	4	121	19/42
Wasson (1929)	IL	9	Algebra, English, Science, Latin	1	4	130	25/40
Davis & Goldzien (1930)	IA	7	History	1	3	140	35/70
Jensen (1930)	CA	H.S.	Algebra	1	6	185	16.3/35.3
Cunningham (1931)	CA	H.S.	Algebra	1	2	83	28/55
Smith (1931)	MN	9	English	1	4	143	20.5/49.5
Whitney (1931), Whitney & Wiley (1932)	CO	H.S.	English, Math	7	22	693	20/40
Brown (1932)	IA	Higher Ed	Psychology, American Government	1	17	679	29.1/55.4
Holy (1932)	OH	H.S.	English	1	12	384	23.5/40
Dalthorp (1934)	SD	H.S.	English	1	2	50	11/39
Hand & Smith (1934)	MN	9	Business	1	3	152	22/105
Ewan (1935)	PA	9–12	Science, French, Latin II, Social Studies	1	8	160	20/40
Eastburn (1936)	AZ	11	American History, English	1	9	360	30/60
Eastburn (1937)	AZ	11	American History, English	1	17	810	30/50

Notes: "English+" denotes the following subjects covered in Stevenson's studies: Reading Comprehension, Reading Rate, Spelling, Language, and Grammar. "H.S." indicates that the grades within high school are not specified. "*n/s*" indicates that the study's author(s) did not provide this information.

This wave of research was motivated in great part by financial pressure on schools, caused by a rise in enrollment at the start of the twentieth century. Census data show that enrollment among children aged 5 to 19 grew from 52 to 72 percent between 1900 and 1930, and, as a result, many school districts had to choose between increasing expenditure—funded almost entirely by local taxes—or increasing class size. The link between class size and school expenditure was especially strong in the early 1900s, with instructional salaries at that time constituting about 80 percent of total expenditure, relative to 45 percent in 1990 (Hanushek and Rivkin, 1997). In addition, construction of new school buildings slowed greatly during World War I, particularly in cities (Burgess, 1920), exacerbating the problem of enrollment growth. Thus, the major question for class size researchers was whether larger classes would have a negative effect on student achievement.

A large fraction of the early field experiments were conducted in the Midwest; of the 24 studies, 13 collected data in one or more of the following states: Iowa, Illinois, Michigan, Minnesota, Ohio, and South Dakota. This regional concentration seems related to institutional factors, as opposed to enrollment or demographic trends. The North Central Association of Colleges and Secondary Schools (the accrediting body for high schools in the Midwest since 1895) imposed class size limits of 25 students, and as many Midwestern high schools started to hit this constraint, they were forced to either hire more teachers or risk losing accreditation. Two of the earliest experimental researchers, Davis (1923) and Stevenson (1923), pointed out that these regulations were not based on any empirical evidence, and sought to bring data to bear on the issue. They were also concerned that these high school regulations were causing greater class size increases in elementary schools, where there was no regulatory constraint and classes were already considerably larger.[3]

This wave of research produced little support for the notion that increasing class size would cause a significant reduction in student achievement. In all but two of the 24 field experiments, the authors concluded that average achievement (or achievement growth) was not significantly reduced in larger classes, and in many instances the students in larger classes outperformed their small class counterparts. According to one contemporaneous literature review (Smith, 1931): "Experiments following each other in rapid succession from 1920 to the present time have proved to us that, so far as the average measurable achievement of our pupils is concerned, our fears for their progress in large classes are unfounded." A subsequent review by Irwin (1937) reached a similar conclusion.

In general, the authors of these early experimental studies made modest claims regarding the policy implications of their findings and were quick to point out the

[3] Data from a survey of cities of 8,000 or more persons (American City Bureau, 1921) indicate median class sizes for high schools and elementary schools of 25 and 38 students, respectively. Today, the median pupil–teacher ratio nationwide is about the same in high schools (15.2) and primary schools (15.6) (U.S. Department of Education, 2008).

need for further research. For example, they were interested in whether class size effects were heterogeneous across students and teachers of different ability levels, whether these effects depended on the homogeneity of students in the classroom, and whether teachers could be trained to work effectively with particularly large or small groups. Nevertheless, the case for small classes was certainly weakened by these studies, and school administrators looking to save money by increasing class size used this evidence to support their case. For example, according to Callahan (1962), the superintendent of the Chicago public schools used the evidence from work by Stevenson and others to place a floor of 20 on class size throughout the city and push for a broad increase in class size. In 1932, the North Central Association responded to the research evidence by eliminating its restrictions and allowing high schools throughout the Midwest to raise class size without the risk of losing accreditation.

Not surprisingly, the use of experimental research to justify increases in class size brought criticism from many professional educators. For example, one complaint was that when researchers focused on student performance on standardized exams, they neglected a host of hard-to-measure but still important educational objectives (Charters, 1929; Hullfish, 1930; Dale, 1931, Pertsch, 1937). Critics also argued that the experimental results might not generalize to real world policy decisions. For example, experimental results might not be informative about overall increases in teaching load if teachers had multiple large sections (Hancock, 1932) and restrictions that some experimental studies placed on teaching methods may have handicapped student performance in the smaller classes (Keener, 1931; Charters, 1933).

In the end, the debate over the potential costs and financial benefits of large classes was short-lived. U.S. birthrates fell during the 1920s and early 1930s, and schools in large cities saw enrollment and class size decline (Smith, 1933; Gates, 1937). As the financial pressure on schools declined, interest in the importance of class size did as well, and class size research all but ceased for the next 15 years.

When the post–World War II baby boom generation began reaching school age, financial pressure on schools brought about a second wave of class size research, but the widespread use of experiments did not resume. The cause of this shift in methodology is impossible to know for certain. Researchers might simply have lost touch after such a long hiatus, although reviews of the class size literature in the 1950s cite earlier experimental work and recognized that "the most productive period from the point of view of research appears to have been the years after the First World War" (Fleming, 1959). A more likely explanation is that this previous work was held in low regard. These experiments lacked statistical rigor—sometimes involving comparisons of mean achievement without any calculation of statistical significance—and many were done on a small scale. In addition, post–World War II researchers had access to large public data sets on educational inputs and outcomes, such as the Coleman Report (Coleman et al., 1966), and those wishing to study class size would no longer have to collect their own data.

Early Experimental Methods

> *In addition to class-size there are several factors which influence the efficiency of instruction such as the intelligence of pupils, the ability of teachers, the spread of ability in the class, the physical equipment of the classroom, the supervision of instruction, and the course of study. It is very evident that any investigation of class-size which does not eliminate or allow for all of these other factors cannot lead to a defensible conclusion.*
>
> Paul Stevenson (1923), Ohio State University

The early field experiments in class size were not conducted by economists: six studies were doctoral dissertations in education, eight were written by academic researchers, and ten were authored by school or school district personnel. As educators, they were quite aware that many factors can influence student achievement, but they possessed only a basic knowledge of statistics. Experiments offered them a relatively simple tool for analysis that could avoid many sources of bias. Their work on class size was part of "a wave of enthusiasm for experimentation in education" (Campbell and Stanley, 1963) that touched upon various topics, such as nontraditional instructional methods and selective grouping of pupils by age or achievement.

In addition, these researchers were interested in results that would be relevant for educational policy, and their experiments were designed to operate within a regular school environment. Outside of the variation in class size, instruction took place under normal conditions. Students attended the same schools, took the same courses, and were taught by the same set of teachers as would have taught them in absence of the experiments.

One key issue in the design of a field experiment in class size is how to ensure that pupil ability and teacher quality are the same in small and large classrooms, and I therefore first discuss how students and teachers were assigned in these early experiments. I then proceed to other important issues for the interpretation of experimental findings: the degree to which the teaching methods were controlled across classrooms; the difference between variation in class size and variation in teaching load (that is, the total number of students or classes assigned to a teacher); and the presence of "Hawthorne effects," in which participants' knowledge of the experimental nature of the study affects their behavior.

Student Assignment

The design of the early class experiments was greatly influenced by the work of William McCall, a professor at Teachers College, Columbia University. In *How to Experiment in Education* (1923), McCall provided practical guidelines for experimental design and measurement, including a lengthy discussion of the assignment of treatment and control groups via randomization and other methods, as well as many other issues that still concern field experimenters today, such as spillover effects between treatment and control subjects. Importantly, he advocated using

measurement to ensure that groups would be "equal in their possibilities for growth in the trait in question" and viewed randomization as "merely an economical substitute for measurement, practicable only where the number of experimental subjects is sufficiently large." However, McCall realized that using measured characteristics to equate groups was not a simple task, and provided a long discussion of methods for equating groups using baseline scores on tests of general ability or subject matter knowledge.

Thus, few of the early field experiments used randomization to assign students to classrooms. While none used a purely random method, such as a lottery, Stevenson (1925a,b) and Hand and Smith (1934) assigned students based on selection at regular intervals from an alphabetically ordered list, a technique suggested by McCall.[4] Interestingly, this was the assignment method used in Project STAR and it has also been used in at least one recent field experiment by economists (Miguel and Kremer, 2004).

Regular selection from an ordered list, however, merits some caution. In one small-scale study (Davis and Goldizen, 1930), students were ranked by prior test scores and those with odd ranks were assigned to large classrooms. While this technique will ensure some degree of similarity across classrooms, it is equivalent to pairing students by prior achievement and assigning the higher scoring member of each pair to the larger class. The authors of this study would have benefited from a careful reading of McCall, who explicitly noted this problem.

In Averill and Mueller (1925) and a subset of the experiments reported by Hudelson (1928) and Judd (1929), a different mechanism was used to generate nearly identical small and large classrooms. First, researchers selected groups of three students that matched closely on a number of characteristics and then assigned one student to a (small) class and two to another (large) class. This setup is similar to "matched pair" designs used by economists to measure discrimination (Neumark, Blank, and Van Nort, 1996; Bertrand and Mullainathan, 2004).

A third type of assignment strategy used in these experiments might be called "match and fill." Here, researchers took a population of students, selected two equal-sized groups that matched closely on a number of characteristics, and placed each group into a separate class. They then added students to one of the classes from the rest of the population, but used only the achievement of the two matched groups to measure the impact of class size. Unfortunately, this assignment method may be subject to bias from peer effects. For example, increasing the size of one class by adding relatively high-achieving peers might negate any negative impact of class size if high-achieving students help their classmates learn. Researchers' understanding of this issue seems mixed. The subset of experiments described by Hudelson (1928) that use this method do not even mention the ability of the

[4] Stevenson does not explain his method in detail in his publications, but his technique is outlined in a report on his ongoing work in the April 2, 1924 issue of the Ohio State Bureau of Educational Research Bulletin. His earlier work (Stevenson 1922, 1923) may have also used this method, but there is no evidence in the texts either to support or reject this notion.

unmatched pupils. In one of the experiments reported by Judd (1929), the large class was filled in with "average students whose IQ's were not taken into consideration." Ewan (1935) claimed to add students in a way that provided a "homogenous environment" in small and large classrooms, but my examination of his data on students' baseline test scores shows this was not always true.

The last and largest category of studies consists of those that did not provide details of an explicit assignment mechanism. Instead, the authors of these studies describe equating student ability across small and large classes, and present summary statistics to demonstrate the similarity of classrooms. One particularly detailed example is given by Whitney and Willey (1932), who state: "The greatest possible care was exercised to insure similarity in paired classes of important factors such as IQ, sex, life age, previous educational achievement, and potential mental age. Judgments of principals and teachers were used in addition to standardized tests so that paired classes would remain approximately equal in ability to do school work." At the other extreme, Tope, Groom, and Beeson (1924) "selected the pupils for each class with no thought of their intelligence" and later checked to be sure the groups were similar. In one instance (Wasson, 1929), baseline intelligence test scores were reported for all students, allowing me to examine assignment in greater detail. Interestingly, despite a vague explanation of the assignment process (students were assigned "so that a number of pupils in the large and small class would have the same [IQ] point-score and the average score for the two classes in each group would be the same"), baseline scores were equal at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the distributions for small and large classes. This provides some confidence that these general assignment procedures may have been successful at assigning similar students to classes of different sizes.

Given the variety of methods used in these early experiments, it is natural to ask what assignment method a modern experiment would use. Most field experiments today use conditional or "stratified" randomization, where participants are first grouped based on a set of characteristics and then a subset of each group is randomly selected into a treatment condition. In the case of class size, institutional constraints require at a minimum that students be stratified at the school and grade level, but further stratification could take place using characteristics such as ethnicity, gender, family income, parental education, and prior achievement. Finer stratification will increase precision (Imbens, King, McKenzie, and Ridder, 2008), and would be highly attractive in an experiment with a small number of classrooms. Note that if the effects of class size are heterogeneous across students, then a highly stratified design also will provide a more "localized" average treatment effect estimate. A stratified design should therefore be selected such that the distribution of classroom characteristics in the experiment is similar to the distribution in the population to whom the results will be generalized.

Teacher Assignment

The second major concern in the design of the early class size experiments was to make teacher quality equivalent in small and large classrooms. McCall (1923)

suggests several ways in which this might be accomplished. The first option is random assignment, but this is likely to have seemed undesirable due to the small number of classrooms in most of these experiments. A second option was to use "objective tests" or "judgments of supervisory officers" to assign teachers of equivalent quality to both types of classrooms. Only one study (Metzner and Berry, 1926) used this method, matching teachers based on evaluations given by their principals and supervisors. Researchers' reluctance to try to equate quality across teachers is understandable. While students' baseline tests are typically powerful predictors of their future achievement, an effective means of gauging teacher quality remains elusive even today, despite a century of research.

In the face of these difficulties, nearly all of these field experiments were designed to compare small and large classes taught by the same teacher.[5] This type of "within-teacher" design has been used in recent educational program evaluations (for example, Barrow, Markman, and Rouse, 2008) and presents several advantages. It is arguably the best option for equalizing teacher ability across treatment and control groups in field experiments containing a small number of teachers. In addition, it can be implemented in circumstances where only one teacher is available to provide instruction in a subject, so long as multiple classes can be formed. Last but not least, by giving teachers both small and large classrooms, it avoids issues of equitable treatment across teachers who participate in the experiment.

At the high school level, instructors typically offer multiple course sections, allowing for a natural and fairly straightforward design of a within-teacher field experiment. In contrast, elementary school teachers typically work with the same group of students for most of the day, which rules out an experimental design where the same individual teaches two elementary school classrooms during the same period of time. The problem was addressed by using McCall's "rotation method," in experiments run by Stevenson (1922, 1923, 1925a), Almack (1922), Holy (1932), and Whitney and Willey (1932). Within each school and grade participating in the study, classrooms that differ in size by some amount N are formed and taught for one semester by two different teachers. At the middle of the year, a representative sample of N students are selected from the large class and moved into the small class, so that sizes of the two classes are interchanged. To estimate the impact of class size on student achievement, the researcher measures achievement growth of the students who remained with the same teacher the entire year and compares the semester in which they were taught in a small group to achievement growth during the semester in which they were taught in a large group. As Stevenson (1923) discusses, this method allows researchers to control for time-specific effects that cause achievement growth to be higher in one semester:

[5] The largest reported study (Davis, 1923) is unfortunately quite vague regarding teacher assignment. It is only stated that small and large classes "were, so far as possible, given identical instruction." However, this was a study of high school students, so the within-teacher design would have been relatively straightforward.

for example, the possibility that "different forms of the achievement tests are not of precisely the same difficulty," or that students may be "out of the school attitude" at the start of the year. It also ensures an equal number of small and large classrooms in each semester, grade, and school.[6]

Teaching Methods and Teaching Load

One pathway by which a smaller class might benefit students is by allowing teachers to adopt alternative methods that raise the quality of instruction. If this is true, then class size variation may have very different effects depending on teachers' reactions to it. This provides a conceptual contrast with the congestion effects model of class size (Lazear, 2001), where a class size reduction provides direct benefits to students even if all other features of the classroom experience are held constant.[7]

The field experiments surveyed here differed in the degree to which teaching methods were allowed to vary between small and large classrooms. Some authors took the idea of an experiment to mean that only one factor—class size—ought to vary, and constrained teaching methods to be the same in both small and large classrooms. Others encouraged the adoption of optimal methods across classrooms of different size. For example, in Davis and Goldizen (1930), "the general plan for the semester, the daily lesson plans, and the verbal instructions in class were as nearly identical in all the sections as it was possible to make them . . . the textbooks, reference materials, and aids to learning were the same for all the sections," while Jensen (1930) left teachers "free to use their own methods and to adapt the work to their pupils . . . If a teacher found one type of instruction more effective with the large than with the small class, she was at liberty to use it."

While the issue of teaching methods is mentioned in a number of studies (often as a direction for future research), Brown (1932) focused on it directly. The author conducted a set of class size experiments with sections of introductory psychology courses at Iowa State Teachers College in 1929. No significant difference was found between performance of students taught in "large experimental" sections and more typically sized "small control" sections, but it was noticed that the courses were taught using the approach that had been typically used in small classes. The following year, more experiments were conducted by dividing students among three types of sections: "large experimental," "small experimental," and

[6] The rotation method could overstate the benefits of smaller classes if the addition of the new students to a classroom—even if the increase takes place over a mid-year break and is fully expected by the teacher—could have a disruptive effect on learning that is distinct from the effect of simply increasing class size. In at least one experiment (Stevenson, 1922, 1923), this may not have been a concern; in Chicago, where the study was conducted, students customarily changed classes in the middle of the year due to semiannual promotion.

[7] In this model, an increase in the number of students in a classroom has a negative impact on achievement through congestion effects. With probability $1 - p$, each student acts in a way that prevents learning from taking place (for example, the student misbehaves and causes the teacher to stop teaching), so that learning only happens during the fraction p^N of class time.

"small control," and methods thought to be more effective with larger classes were used in the experimental classes. In every case, the large classes significantly outperformed both comparison groups. These findings were based on a small sample of 14 course sections and 571 students, but they suggest that optimal methods of teaching are important to incorporate into thinking about the impact of class size.

Restrictions on variation in teaching methods in class size experiments present a tradeoff between practical knowledge and deeper insight, which is analogous to a greater discussion in economics of structural and reduced-form empirical methods. Allowing for full adjustment by teachers, students, and others in response to experimental variation in class size may be of more immediate use for policy decisions, even if it cannot help us definitively answer the deeper question of why class size might matter. In contrast, constraining all other factors may deepen our understanding of the role of class size in the educational production function, but may generate wildly inaccurate estimates of the expected impact of changes in class size policies.

A similar issue of interpretation arises when we consider the possible importance of teaching load. It is reasonable to believe that the quality of educational provision will bear some relation to the total number of classrooms and/or students with which the teacher has to work. Consider a teacher with two sections of high school math, sections A and B, each with 30 students. If we add students to section A, how much does this affect achievement in section B? If we reduce class size to 20 by pulling students from sections A and B and creating a third section of equal size, are students better off? The answers to these questions are important for interpreting the results of class size experiments. The vast majority of the early field experiments did not involve large reductions in teaching load. Teachers were given two sections, one smaller than average and one larger, leaving their overall load relatively stable and eliminating the need to hire additional teachers in order to conduct the experiment. As with teaching methods, constraining teaching load in this way provides a sharper focus on the marginal effects of class size variation, but is not necessarily informative about the effects of a general increase in the size of all sections assigned to a teacher.[8]

Hawthorne Effects

One general concern with field experiments is the possibility that when participants know they are being studied, they may behave differently. The current

[8] Only one field experiment from this era (Holy, 1932) examined a general shift in both class size and teaching load. In the fall of the school year 1929–1930, two teachers "rated as equally efficient by the Superintendent's office" were given very different teaching loads in high school English: six classes of 24 students on average versus six classes of 40 students on average. In the second semester, the classes were maintained but the teachers were switched, and students' test score gains were calculated within both semesters using four examinations. The author reports no statistically significant difference in gains between the large and small classes but does not provide any data that could be used for replication.

term for this phenomenon derives its name from experiments conducted at a factory in the Hawthorne suburb of Chicago in 1924, but the issue was discussed at length in McCall's (1923) book on experimentation in education. McCall discussed several ways in which experimental results might be influenced by "conscious or unconscious manifestation of bias" on the part of subjects, teachers, and experimenters themselves, and recommended keeping "pupils in ignorance of the nature of the experimental factors and, if possible, of the fact that an experiment is in progress."

Comments related to this issue are rare in the early class size studies, and the few exceptions suggest that researchers took different views of the subject. Almack (1922) reports that "pupils were not informed that an experiment was under way, and there is no evidence that they became aware of it," and Cunningham (1931) states that students "were told that the groups had been formed for experimental purposes, though the precise nature of the experiment was not disclosed to them." In contrast, Jensen (1930) informed teachers and students that "the objective of the experiment was to see which class could achieve the greatest mastery of the subject," and that among students "the rivalry in some of them, at least, was very spirited at times." Nevertheless, it is reasonable to think that Hawthorne effects were not responsible for the general conclusion from these field experiments that larger classes were not detrimental to students. After all, teachers and students (often surveyed by experimenters) typically preferred smaller classes and thus had an incentive to act in ways that would bias experiments in their favor.

Other Concerns

Modern researchers carrying out field experiments would raise some other issues that were not addressed in this earlier literature. First, these researchers almost never mention attrition bias—the possibility that results may be affected by students leaving the sample in a nonrandom fashion. The single instance in which I found mention of the issue is in Stevenson (1923), who, in order to ensure that "classes were maintained intact as far as possible throughout the year," arranged that "promotion standards were waived and that all pupils were advanced at the end of the first semester." So, gauging the extent of attrition in the early class size studies and how it might have affected authors' conclusions is unfortunately impossible to do with the data they provide.

A second concern that present-day researchers would take into account is that students in the same classroom should not be considered as independent observations when making inferences about statistical significance. Class size does not vary within a classroom, and various common factors affecting all student outcomes within a classroom (like teacher quality) will lead traditional estimates of standard errors to be biased downward. This problem is typically handled by allowing for clustering in the calculation of standard errors (Moulton, 1986), but this solution is infeasible when there are a small number of clusters. An alternative in these early field experiments would have been to treat each classroom as a single observation or use a two-step procedure such as that proposed by Donald and Lang (2007) to

incorporate control variables that vary within classrooms. Ultimately, this issue makes it difficult to obtain a precise estimate of the impact of an educational policy with just a few classrooms, and perhaps the smallest of these early field experiments would have been abandoned had this been well-understood at the time.

Early Experiments and Project STAR

The consensus among early experimental researchers was that increasing class size does not inhibit student learning. Although some authors provided careful documentation of their data and analysis, many reported only a smattering of class means, which leaves no way of using statistical analysis to confirm their findings, and it is therefore impossible to combine the results of these 24 field experiments into a single set of estimates.[9] However, to provide some confirmation of these early findings, I reanalyze the data reported in Stevenson (1922, 1923, 1925a), which was based on field studies in elementary schools. While student-level data are not available, Stevenson reported average test score gains at the classroom-semester cell level in terms of student-level standard deviations. Stevenson's work was the most widely cited and was, in my view, of high quality. These experiments contained relatively large samples, together totaling 78 classrooms and over 2,000 students in grades two, five, and seven. Stevenson also used the same design and measurement methods in both experiments, making it possible to aggregate their results.

I conduct a simple retrospective analysis, running a regression of the average change in classroom achievement on an indicator variable for being in a large class, pooling data across grades and tested subjects.[10] On average, large classes contained 43.6 students, compared with 28.2 in small classes. I estimate that being in a large class was associated with an average *increase* in achievement growth of 0.010 standard deviations, with a standard error of 0.055. While this estimate is less precise than might be desired, it is sufficient to rule out negative effects of increased class size on the order of 0.10 standard deviations—about half of the class size effect found in Project STAR.[11] Given these results, one might concur with Stevenson that, at least within the context he examined, "it is evident that results taken from many large and small classes in grade [sic] II, V, and VII do not favor small classes."

There are many possible reasons why the conclusion drawn by Stevenson and his contemporaries—that students do not learn more in smaller classes—contrasts with the results of Tennessee's Project STAR. One potential explanation is that the

[9] Glass and Smith (1978) include a subset of these field experiments (and some nonexperimental studies from this period) in their meta-analysis.

[10] More details on this regression analysis and additional biographical detail about Stevenson are available in an online appendix to this paper available at ⟨http://www.e-jep.org⟩.

[11] Note that the measure of standard deviation here is based on Stevenson's samples within each study. If the results were expressed in standard deviations of a broader, nationally representative population (as were the exams taken by students in Project STAR), the estimate and its standard error would be smaller in magnitude.

earlier studies had serious methodological shortcomings. Although I have argued that these experiments were usually designed carefully, without full access to their data it is impossible to reproduce their results using modern econometric methods.

However, even if we take their results as given, it is worth considering the differences in both experimental design and setting that could explain the two sets of findings. One obvious difference is that Project STAR focused on students in early childhood (grades K–3), while few of the earlier experiments included students from this age group. Indeed, none of these studies included kindergarten students, where the benefits of class size in Project STAR were arguably the clearest.

Another issue is whether the variation in class size in these early experiments was sufficient to generate meaningful differences in student achievement. If we consider proportional variation, it is hard to argue that the experimental changes in class size were small; large classes were typically 50 to 100 percent larger than small classes. Still, the focus of these early experiments was on the detrimental effects of large classes, and comparisons were typically made between small classes of 20–35 students and large classes of 35 to 50.[12] It is quite plausible that variation along this part of the class size distribution may have a negligible impact on achievement while variation along the lower end of the distribution is far more important. A close review of the early experimental studies provides mixed support for this hypothesis. While the two authors presenting results in favor of class size reduction examined small classes of less than 20 students, three other experiments using similarly sized classes did not find positive effects of reduced class size.

Another potentially important difference is the duration of the experiments. The class size reductions in Project STAR lasted four years, and the early experiments may not have lasted long enough for the effects of class size to emerge. This seems unlikely, however, given that significant effects of class size in Project STAR were found after one year. While one of the early experiments lasted only nine weeks, twelve lasted for one semester, and eleven lasted a full school year. While students in the early experiments were not tracked to examine possible long-term effects, researchers presumably saw little reason for such efforts given the short-run results.

One more concern regarding research design is that, unlike Project STAR, nearly all of the early experiments were based on variation among classes taught by the same teacher. Even if teachers were given free reign to adjust their methods across classes, whether they did so is an open question. If the fixed costs of class preparation are large, then a teacher with classes of different sizes might optimally

[12] See Table 1 for information on the size of small and large classes in these experiments. There are two notable outliers. In light of prior results showing little effect of class size on student achievement, Hand and Smith (1934) sought to undertake an experiment where the definition of a large class underwent a "rather drastic revision." To that end, they compared two classes of 22 and 25 pupils with a large class of 105. Hudelson (1928) presents the results of a number of experiments in various departments at the University of Minnesota, one of which involved two small classes of twelve and two large classes of 138 and 148 students.

choose a single method that could be suboptimal for either section taught alone.[13] Another difference from Project STAR is the within-teacher design of the early experiments, which does not alter total teaching load. It is possible that teaching load is a more important determinant of the quality of instruction than class size. In elementary grades, there is typically only one class per teacher, so the effects of class size and teaching load cannot be separately identified.

A final possibility is that the relationship between class size and student achievement may have changed over the twentieth century. Many important transformations occurred in the U.S. educational system during this period: the spread of high school education to the vast majority of the population, the shift away from local sources of revenue, the consolidation of school districts, the unionization of teachers, racial desegregation, and the proliferation of special education programs for students with disabilities. Yet in many ways, the production of education has also remained remarkably consistent. The length of the school year and attendance rates changed little after 1930 (National Center for Education Statistics, 1993), and the way in which students and teachers are organized within schools is much the same as in the decades prior to World War II (Cuban, 1993).

Even if we cannot explain why the earlier results differed from Project STAR, one lesson provided by the early experimental literature is that no single study can provide a clear and comprehensive understanding of such a complex issue. While few would deny the internal validity and impressive results of Project STAR, many economists would also agree with Mishel and Rothstein (2002) that "enthusiasm for it has been somewhat excessive." Given the importance attached to experimental evidence, it is somewhat puzzling that no other large-scale class size experiment been conducted in close to 25 years.

Whither Class Size Experiments?

Interest in the issue of class size has not waned among economists who study education. However, rather than conduct field experiments, economists have largely focused on using instrumental variables to obtain causal estimates in nonexperimental settings. One popular strategy, first used by Angrist and Lavy (1999) and Hoxby (2000), has been to take advantage of a discontinuous relationship between enrollment and class size created by class size limits. While this method has since been applied in various studies spanning many countries, it has produced mixed evidence; some studies find positive significant effects of reduced class size and others find fairly precise estimates indicating no relationship between class size and achievement. In addition,

[13] On the other hand, it is not clear that relying on random assignment and between-teacher variation would lead to greater adjustment of methods if a teacher does not expect the class size variation to persist for a long period of time. Indeed, one puzzling finding in the literature on Project STAR is that measures of teaching practices did not differ significantly between small and large classes (Evertson and Randolph, 1989).

Urquiola and Verhoogen (forthcoming) show that this line of inquiry may produce biased results because students can sort in ways that generate discontinuities in student characteristics around class size limits. An alternative instrumental variables strategy, developed by Hoxby (2000), is to examine the small, idiosyncratic fluctuations in cohort size that occur within a particular school and grade level over time. This work has produced little support for reduced class size (Hoxby, 2000; Wößmann and West, 2006; Leuven, Osterbeek, and Rønning, 2008). However, identification based on idiosyncratic variation in cohort size may be of limited use in evaluating the potential benefits of major class size reduction policies.

The continued debate over class size suggests reconsideration of the field experiment methodology. However, there are a number of obstacles to the organization of a large-scale class size experiment, perhaps most importantly the cost of hiring teachers. The salary expense for 130 additional teachers—roughly the number needed each year in Project STAR—would be about $6 million today, well outside the typical range for federal research grants. In addition to financial costs, it may also be difficult to gain the support of the various interested parties—parents, teachers, principals, and others—whose cooperation would be needed to implement a randomized class size intervention in an effective manner. Even in the case of Project STAR, which was supported and funded by the Tennessee state government, only about one in five eligible schools volunteered to participate. Thus, it is unlikely that researchers could implement the Project STAR methodology without a large school district or state government willing to provide financial and political support.

Nevertheless, less costly research designs are possible, such as increasing the size of some classes in order to reduce others. Elementary schools with 100 students per grade could organize one class of 16 and three classes of 28. High schools could be randomly selected to decrease class size for either (say) Algebra or Geometry, while increasing class size in the other subject in order to maintain teaching loads. Increasing the size of some classes in order to decrease others may make gaining the cooperation of teachers more difficult, but one might address this in a multiyear experiment by rotating teachers across large and small classrooms in a randomized order. Field experiments could also follow their historical roots and use within-teacher variation in class size, removing the need to hire new teachers. Finally, it is worth noting that variation in class size can also be embedded within a larger experiment to estimate causal impacts of multiple policies and policy interactions. For example, Duflo, Dupas, and Kremer (2007) study class size reduction, peer effects, and teacher incentives within the context of a single experiment.

While there are limits to what experiments reveal about economic behavior (see Heckman and Smith, 1995, for a discussion), this method has grown in stature among economists (Harrison and List, 2004; Levitt and List, 2007), and is widely viewed by educational policy makers as a "gold standard" for empirical research.[14] Thus, the potential impact of experimental evidence on class size reduction is considerable.

[14] For example, the Education Sciences Reform Act of 2002, which provides federal funding for research in education, explicitly distinguishes experimental study designs as of the greatest merit.

Indeed, in addition to triggering academic publications, the results of Project STAR have motivated states like California and Florida to spend billions of dollars on reductions in class size in ways that are considerably different than the experimental intervention in Tennessee. In particular, these class size reductions are usually modest and have sometimes been directed outside of the early childhood grades. With credible information about the likely benefits of these class size reduction policies, state governments and school districts can make better decisions about whether they are worth the cost.

■ *This paper has benefited greatly from many thoughtful comments and suggestions given by David Autor, Ray Fisman, Eric Hanushek, James Hines, Caroline Hoxby, Brian Jacob, Alan Krueger, Leigh Linden, Andrei Shleifer, Doug Staiger, Jeremy Stein, Timothy Taylor, and Miguel Urquiola. Miya Hirabayashi Virbalas provided excellent research assistance.*

References

Almack, John C. 1922. *The Adaptation of the School Building to a Program of Educational Efficiency.* Unpublished doctoral dissertation, Stanford University.

American City Bureau. 1921. *Know and Help Your Schools, Second Report.* New York: American City Bureau.

Angrist, Joshua D., and Victor C. Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics,* 114(2): 533–75.

Averill, Lawrence A., and Alfred D. Mueller. 1925. "Size of Class and Reading Efficiency." *The Elementary School Journal,* 25(9): 682–91.

Bain, Helen P., and C. M. Achilles. 1986. "Interesting Developments on Class Size." *Phi Delta Kappan,* 67(9): 662–65.

Barrow, Lisa, Lisa Markman, and Cecilia Rouse. 2008. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." NBER Working Paper 14240.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review,* 94(4): 991–1013.

Brown, Albert E. 1932. "The Effectiveness of Large Classes at the College Level: An Experimental Study Involving the Size Variable and Size-Procedure Variable." *University of Iowa Studies in Education,* 7(3): 1–66.

Burgess, Warren R. 1920. *Trends of School Costs.* New York: Russell Sage Foundation.

Callahan, Raymond E. 1962. *Education and the Cult of Efficiency; A Study of the Social Forces That Have Shaped the Administration of the Public Schools.* Chicago: University of Chicago Press.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: R. McNally.

Charters, Werrett W. 1929. "Larger Classes." *Educational Research Bulletin,* 8(12): 276.

Charters, Werrett W. 1933. "Co-ordination of Instruction." *The Journal of Higher Education,* 4(3): 125–130.

Coleman, James S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. L. York. 1966. *Equality of Educational Opportunity.* Washington, DC: U.S. Government Printing Office.

Cuban, Larry. 1993. *How Teachers Taught: Constancy and Change in American Classrooms 1880–1990.* New York: Teachers College Press.

Cunningham, Margaret S. 1931. An Experiment in Class Size in B9 Algebra. *California Quarterly of Secondary Education,* October, vol. 7, pp. 19–33.

Dale, Edgar. 1931. "Editorial Comment." *Educational Research Bulletin,* 10(9): 243–44.

Dalthorp, Charles J. 1934. "An Experiment with a Large and Small Class in English Composition." *High School Teacher,* February, vol 10, pp. 51.

Davis, Calvin O. 1923. "The Size of Classes and the Teaching Load in the High Schools Accredited by the North Central Association." *The School Review,* 31(6): 412–29.

Davis, Everett, and Goldizen, Mae. 1930. "A Study of Class Size in Junior High School History." *The School Review,* 38(5): 360–67.

Donald, Stephen G., and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics,* 89(2): 221–33.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2007. "Peer Effects, Pupil–Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya." Unpublished manuscript.

Eastburn, Lacey A. 1936. "The Relative Efficiency of Instruction in Large and Small Classes on Three Ability Levels." Doctoral dissertation, Stanford University.

Eastburn, Lacey A. 1937. "Report of Class Size Investigations in the Phoenix Union High School, 1933–34 to 1935–36." *Journal of Educational Research,* 31(2): 107–117.

Evertson, Carolyn M., and Catherine H. Randolph. 1989. Teaching Practices and Class Size: A New Look at an Old Issue. *Peabody Journal of Education,* 67(1): 85–105.

Ewan, Stacy N. 1935. "The Relation of Class Size and Selected Teaching Methods to Pupil Achievement." Unpublished doctoral dissertation, University of Pennsylvania.

Fleming, C. M. 1959. Class Size as a Variable in the Teaching Situation. *Educational Research,* 1(2): 35–48.

Gates, Arthur I. 1937. "The Problem of Class Size in the Elementary Grades." *The Elementary School Journal,* 37(6): 405–7.

Glass, Gene V., and Mary L. Smith. 1978. *Meta-analysis of Research on Class Size and Achievement.* San Francisco, CA: Far West Laboratory.

Hancock, J. Leonard. 1932. "Shall the Sky be the Limit?" *The Journal of Higher Education,* 3(6), 294–296.

Hand, Harold C., and J. W. Smith. 1934. "Effectiveness of Instruction in a Class Group of One Hundred Pupils." *The School Review,* 42(10): 751–54.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature,* 24(3): 1141–77.

Hanushek, Eric A., and Steven G. Rivkin. 1997. "Understanding the Twentieth-century Growth in U.S. School Spending." *The Journal of Human Resources,* 32(1): 35–68.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature,* 42(4): 1009–1055.

Heckman, Jeffrey A., and James H. Smith. 1995. "Assessing the Case for Social Experiments." *The Journal of Economic Perspectives,* 9(2), 85–110.

Hedges, Larry V., Richard D. Laine, and Rob Greenwald. 1994. An Exchange: Part I: Does Money Matter? A Meta-analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher,* 23(5): 5–11.

Holy, Thomas C. 1932. Effect of Total Teaching Load on the Efficiency of Instruction in High School English. *National Education Association Proceedings,* 4(3): 365–66.

Hoxby, Caroline M. 2000. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics,* 115(4): 1239–85.

Hudelson, Earl. 1928. *Class Size at the College Level.* Minneapolis: University of Minnesota Press.

Hullfish, H. Gordon. 1930. "A Problem in College Teaching." *The Journal of Higher Education,* 1(5): 261–68.

Imbens, Guido W., Gary King, David McKenzie, and Geert Ridder. 2008. On the Finite Sample Benefits of Stratification in Randomized Experiments. Unpublished paper.

Irwin, Manley E. 1937. "Size of Class and Teaching Load." *Review of Educational Research,* 7(3): 276–83.

Jensen, Milton B. 1930. The Influence of Class Size upon Pupil Accomplishment in High School Algebra. *Journal of Educational Research,* 21(5): 337–56.

Judd, Charles H. 1929. Report of the Consultative Committee. *Bulletin: Department of Secondary School Principals,* March, vol. 25, pp. 49–61.

Keener, E. E. 1931. "What Size Class." *The Elementary School Journal,* 32(2): 144–46.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics,* 114(2): 497–532.

Krueger, Alan B. 2003. "Economic Considerations and Class Size." *The Economic Journal,* 113(485): F34–F63.

Krueger, Alan B., and Diane M. Whitmore. 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(468): 1–28.

Lazear, Edward P. 2001. "Educational Production." *The Quarterly Journal of Economics,* 116(3): 777–803.

Leuven, Edwin, Hessel Oosterbeek, and Marte Rønning. 2008. "Quasi-Experimental Estimates of the Effect of Class Size on Achievement in Norway." IZA Working Paper 3474.

Levitt, Steven, and John A. List. 2007. "What Do Laboratory Experiments Measuring Preferences Tell Us about the Real World?" *Journal of Economic Perspectives,* 21(2): 153–74.

McCall, William A. 1923. *How to Experiment in Education.* New York: The Macmillan Company.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica,* 72(1): 159–217.

Mishel Lawrence, and Richard Rothstein, eds. 2002. *The Class Size Debate.* Washington, DC: Economic Policy Institute.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics,* 32(3): 385–97

Metzner, Alice B, and Charles S. Berry. 1926. "Size of Class for Mentally Retarded Children." *Training School Bulletin,* October, vol. 23, pp. 241–51.

National Center for Educational Statistics. 1993. *120 Years of American Education: A Statistical Portrait.* Edited by Thomas Snyder. U.S. Department of Education.

Neumark, David, Roy J. Bank, and Kyle D. Van Nort. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *The Quarterly Journal of Economics,* 111(3): 915–41.

Pertsch, C. Frederick. 1937. "Research and the Problem of Optimum Class Size." In *Yearbook of the New York Society for the Experimental Study of Education,* pp. 23–27.

Smith, Dora V. 1931. "Experiments in Handling Larger Classes." *The English Journal,* 20(5): 371–78.

Smith, Dora V. 1933. "Vital Factors in the Present Situation in Class Size." *The English Journal,* 22(5): 366–74.

Stevenson, Paul R. 1922. "Relation of Size of Class to School Efficiency." *University of Illinois Bulletin,* 14(45): 1–39.

Stevenson, Paul R. 1923. *Smaller Classes or Larger; Relation of Class-Size to the Efficiency of Teaching.* Bloomington, IL: Public School Publishing Company.

Stevenson, Paul R. 1925a. "Class-size in the Elementary School." Bureau of Educational Research Monographs, No. 3.

Stevenson, Paul R. 1925b. "More Evidence Concerning Small and Large Classes." *Educational Research Bulletin,* 4(11): 231–33.

Tope, Richard E., Emma Groom, and Marvin F. Beeson. 1924. "Size of Class and School Efficiency." *Journal of Educational Research,* 9(2): 126–32.

Urquiola, Miguel, and Eric A. Verhoogen. Forthcoming. "Class Size and Sorting in Market Equilibrium: Theory and Evidence." *American Economic Review.*

U.S. Department of Education, National Center for Education Statistics. 2008. Common Core of Data (CCD) 2005–2006. Retrieved August 27, 2008, from http://nces.ed.gov/ccd/.

Wasson, William H. 1929. "A Controlled Experiment in the Size of Classes." Unpublished master's thesis, University of Chicago.

Whitney, Frederick L. 1931. "The Trinity-Pueblo Class Size Experiment in the Primary School." Department of Educational Research Study No. 113, Colorado State Teachers College.

Whitney, Frederick L., and Gilbert S. Willey. 1932. Advantages of Small Classes. *School Executives Magazine: The American Educational Digest,* August, vol. 51, pp. 504–6.

Wößmann, Ludger, and Martin R. West. 2006. Class-Size Effects in School Systems around the World: Evidence from Between-Grade Variation in TIMSS. *European Economic Review,* 50(3): 695–736.

Word, Elizabeth, John Johnston, Helen P. Bain, B. DeWayne Fulton, Jayne B. Zaharias, Martha N. Lintz, Charles M. Achilles, John Folger, and Carolyn Breda. 1990. *Student/Teacher Achievement Ratio (STAR): Tennessee's K–3 Class Size Study, Final Report.* Nashville, TN: Tennessee State Department of Education.

This article has been cited by:

1. Torberg Falch, Astrid Marie Jorde Sandsør, Bjarne Strøm. 2017. Do Smaller Classes Always Improve Students' Long-run Outcomes?. *Oxford Bulletin of Economics and Statistics* **79**:5, 654-688. [Crossref]
2. Joseph Han, Keunkwan Ryu. 2017. Effects of class size reduction in upper grades: Evidence from Seoul, Korea. *Economics of Education Review* **60**, 68-85. [Crossref]
3. Miguel Urquiola. 2015. Progress and challenges in achieving an evidence-based education policy in Latin America and the Caribbean. *Latin American Economic Review* **24**:1. . [Crossref]
4. Steven G. Dieterle. 2015. Class-size reduction policies and the quality of entering teachers. *Labour Economics* **36**, 35-47. [Crossref]
5. Gerald Eisenkopf, Zohal Hessami, Urs Fischbacher, Heinrich W. Ursprung. 2015. Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland. *Journal of Economic Behavior & Organization* **115**, 123-143. [Crossref]
6. Esther Duflo, Pascaline Dupas, Michael Kremer. 2015. School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics* **123**, 92-110. [Crossref]
7. A. Leigh. 2013. The Economics and Politics of Teacher Merit Pay. *CESifo Economic Studies* **59**:1, 1-33. [Crossref]
8. Matthew M. Chingos. 2013. Class Size and Student Outcomes: Research and Policy Implications. *Journal of Policy Analysis and Management* **32**:2, 411-438. [Crossref]
9. Núria Rodríguez-Planas. 2012. Mentoring, educational services, and incentives to learn: What do we know about them?. *Evaluation and Program Planning* **35**:4, 481-490. [Crossref]
10. Matthew M. Chingos. 2012. The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review* **31**:5, 543-562. [Crossref]
11. Keith Morrison, Greetje van der Werf. 2012. Editorial. *Educational Research and Evaluation* **18**:7, 617-619. [Crossref]
12. T. S. Dee, M. R. West. 2011. The Non-Cognitive Returns to Class Size. *Educational Evaluation and Policy Analysis* **33**:1, 23-46. [Crossref]
13. John A. List, Imran Rasul. Field Experiments in Labor Economics 103-228. [Crossref]