Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size

FREDERICK MOSTELLER
Harvard University and the American Academy of Arts and Sciences

RICHARD J. LIGHT
JASON A. SACHS
Harvard University

In this article, Frederick Mosteller, Richard Light, and Jason Sachs explore the nature of the empirical evidence that can inform school leaders' key decisions about how to organize students within schools: Should students be placed in heterogeneous classes or tracked classes? What is the impact of class size on student learning? How does it vary? Since tracking (or skill grouping, as the authors prefer to call it) is widely used in U.S. schools, the authors expected to find a wealth of evidence to support the efficacy of the practice. Surprisingly, they found only a handful of well-designed studies exploring the academic benefits of tracking, and of these, the results were equivocal. With regard to class size, the authors describe the Tennessee class size study, using it to illustrate that large, long-term, randomized controlled field trials can be carried out successfully in education. The Tennessee study demonstrates convincingly that student achievement is better supported in smaller classes in grades K-3, and that this enhanced achievement continues when the students move to regular-size classes in the fourth grade and beyond. The authors suggest in conclusion that education would benefit from a commitment to sustained inquiry through well-designed, randomized controlled field trials of education innovations. Such sustained inquiry could provide a source of solid evidence on which educators could base their decisions about how to organize and support student learning in classes and schools.

Snapshot of the U.S. Education System

U.S. schools form a vast, expensive, and complicated enterprise. Each school day, the United States spends $1.5 billion on its schools. In 1994, U.S. public schools spent a total of $285 billion on students in kindergarten through twelfth

Harvard Educational Review Vol. 66 No. 4 Winter 1996

grade, offering an average of 190 days of instruction. In the same year, religious and independent schools spent an additional $30 billion. Throughout the United States, 15,000 school districts employ more than 2.5 million teachers, who teach more than 44 million students in 84,000 schools.

Organizing Students and Choosing Class Sizes

Each school leader must make critical decisions about how to organize students within his or her school. Do students learn better when they are grouped into classes in special ways? Are students most productive when, after age grouping, they are divided into classrooms randomly? Or might some systematic grouping process serve students better? A second major decision — determining size of classes — depends on what is known about the impact of different class sizes. Should all classes be of similar size? Does learning take place more effectively when certain classes are especially small and others are larger?

This article focuses on these two pervasive issues: organizing students into classes, and the impact of class size on students' learning. We embark on these topics because they are clearly important to school leaders, to teachers, and especially to parents with children in schools. In the first part of the article we review skill grouping. In the second part we review class size. Both parts include a detailed summary of the main findings — on the impact of skill grouping on student learning in part one, and about the impact on learning of choosing certain class sizes in part two. Part three uses information from the first two parts to explore the need for sustained inquiry to improve practice in education.

Overview of Findings

Our exploration reveals that too little sustained evaluation of current practice and innovations is now being carried out. As a result, policymakers do not have the information needed to make wise decisions within our education system. For example, the studies of skill grouping that we review in this article are nearly all small-scale and short-term, making it difficult for policymakers to draw conclusions. Policymakers need to be able to generalize results to diverse populations of children, and to have confidence that their inferences are not based on idiosyncratic results from a particular sample. Medicine has learned a great deal from large-scale studies of this kind over the past fifty years; education should benefit from additional large experiments as well. Thus, our major conclusion is that leaders in education need to initiate more large-scale, long-term evaluations.

U.S. education does not lack innovations; rather, it lacks careful, long-term evaluations of their performance. In order to be evaluated well, an intervention must be implemented in enough depth so that it is well defined. Teachers must develop sufficient experience to actually deliver it. Then, after initial evaluation, one would expect adjustments and improvements, followed by further evaluation. Our impression is that this process does not often take place in education.

Instead, innovations are introduced, but frequently without sustained evaluation.

The evidence on the impact of grouping students by skill level is limited, offering too little firm guidance, dramatizing the need for more exploration and evaluation. Our review of skill grouping turns up only a few rigorous experiments. We salute the small group of practitioners who are experimenting with different ways of organizing students and delivering instruction, and who are also systematically evaluating the impact of these different options.

A series of exemplary investigations carried out in Tennessee offers a sharp contrast between studies of class size and the skill grouping studies. Results from the Tennessee studies inform policymakers how different class sizes actually affect students' learning. Relying on these results, school leaders and teachers can confidently make certain decisions involving the trade-offs between larger versus smaller class size. Our review of class size also tells policymakers that it is possible to do excellent, rigorous research on a large scale — in many schools, with many children, over a long time, using a well-designed plan.

These two reviews teach us that the payoff from buckling down to implement a well-designed field study can be high. To meet growing demands for excellence in education, we need more evidence about what works well and what does not. In sum, our review of skill grouping illustrates the need for more exploration and evaluation; our review of class size offers a compelling illustration that large-scale field experiments in schools actually can be done.

SKILL GROUPING

The National Educational Longitudinal Survey (NELS) of 1988, described by Chubb and Moe (1992), is a carefully designed ongoing survey of 25,000 students in nearly 1,000 schools. Initially, it followed students for three years, from eighth grade in 1988 through tenth grade in 1990. This large-scale survey examines the use of ability grouping, or, as we prefer to call it, skill grouping. We prefer the expression "skill grouping" rather than "ability grouping" because the latter suggests a sense of permanence in a quality that we believe might be modified by education, training, and practice. Skill grouping, on the other hand, suggests that students sharing a similar current skill level are grouped together for purposes of instruction.

The NELS survey reveals that approximately 86 percent of public school students in U.S. middle and high schools are currently placed in skill-grouped classes for mathematics instruction. Independent schools implement this practice at a slightly lower rate of 71 percent. These numbers tell us that an overwhelming majority of U.S. students are skill grouped for math instruction.

Four Kinds of Skill Grouping for School Instruction

What methods of skill grouping are now used to allocate students among classes? The common starting point for most school systems is age grouping by grades, because, on average, older children have higher skill levels than younger ones. A few schools, such as those in Wellesley, Massachusetts, are initiating programs to bring children of different ages into the same classroom (D. B. Pillemer, personal communication, 1995). For our work, we accept initial age grouping without review. Besides age grouping, four main forms of skill grouping describe the practices now in use.

Whole-Class or Mixed Grouping: Heterogeneous Grouping within Grades

Whole-class instruction is now used in many schools, in which all students in a grade are taught as a group. If the grade includes too many children to be taught in one classroom, the students are separated into groups so that each group or each classroom represents the whole spectrum of students' skills. This grouping produces heterogeneous classes, because the skill levels of the children within each class usually vary considerably. Such whole-class instruction, sometimes called mixed grouping, often serves as a control group in experimental studies that assess the effectiveness of other forms of skill grouping.

Between Class Grouping or XYZ Skill Grouping: Homogenous Grouping within Grades

A second method of allocating students is called between-class grouping, or XYZ grouping. In this method, students in a grade are stratified, usually into two or three levels of skill, such as high, medium, and low. This type of grouping is implemented by using prior achievement in the subject being taught, or by performance on a general aptitude test, or it may be based on some overall rating by the teacher. For convenience we speak of three levels, where the high-skill, medium-skill, and low-skill students are taught in separate classes.

In most studies of XYZ grouping, only slight adaptations of the curriculum to the skill level of students in different classes have occurred. Often the investigator mentions the desirability of such adaptation and regrets that the actual study either did not use adaptation or did not produce information about such adjustments. In some school systems, courses are constructed especially for extraordinarily gifted children or for children with special needs, although we do not review such courses in this article.

Cross-Grade Grouping or the Joplin Plan: Homogenous Grouping across Grades

A third, less common, but much talked-about plan is based upon cross-grade grouping, sometimes referred to as the Joplin Plan. An article, "Johnny Can Read in Joplin," on the use of this plan in Joplin, Missouri, appeared in the *Saturday Evening Post* of October 26, 1957. Let us illustrate with an example from grades 4, 5, and 6. For the purpose of teaching reading, teachers might abandon the distinction between these three grades and focus instead on each student's skill level for reading. Among these three grades of students, skill levels can range widely — perhaps from reading level grade one through reading level grade nine. To handle this great variation, cross-grade grouping might form classes for nine different levels of reading skill. When working on reading, each student joins other students who have the same skill level that he or she has achieved, regardless of original grade level (4, 5, or 6). Students at the same reading level all work on the same material. When reading class is over, students return to their original grades. Having mastered the reading material at one level, the student immediately moves up to the next level of reading skill.

Clearly this approach differs from the XYZ grouping because the material being taught is matched to each student's accomplishments, whether at the level of comic books or Shakespeare's plays. Students working at different levels study different materials fitted to their skill level. When this method is applied to more than one topic, some students will be at different levels in different topics, such as reading and arithmetic.

Within-Class Grouping: Homogenous Grouping within Classes

A fourth way to sort students is within-class grouping. Here, the teacher of a whole class sorts the students into subgroups within the class based on their skill levels, often using three levels, as in XYZ grouping. But the key distinction is that all three subgroups of students stay in the same classroom. While the teacher teaches one skill subgroup a new lesson in arithmetic, for example, other skill subgroups work on arithmetic assignments given the day before. The teacher gives short lessons to each subgroup separately. After all three subgroups have worked on arithmetic assignments, the teacher may have a little time to discuss the same or new work with the whole class. Subgroups within a classroom may have somewhat different assignments, and their goals may not be identical.

Other Forms of Grouping

Beyond these four methods of skill grouping, some other special teaching innovations are currently being explored. For example, in a variation of within-class grouping, special groups are formed called teams (Slavin, Madden, & Leavey, 1984; Slavin, 1995). Each team is likely to be a cross-section of the whole class, because the teams should be approximately equivalent to one another in skill level. Each team has special responsibilities in carrying out the education of its members. They help to instruct one another, and the team checks its own members' work, keeps track of completed assignments, and keeps records of scores on tests and of other activities.

These examples give just a taste of the many possibilities for skill grouping. How can a school leader make a wise choice? We review the experimental studies that compare whole-class instruction with XYZ grouping, the Joplin Plan, and within-class grouping later in this article.

The Controversy about Skill Grouping

Skill grouping generates vigorous controversy. Oakes (1986) has written that skill grouping or ability grouping (or "tracking," as some educators call this practice) inevitably separates not only academically stronger from less strong students, but also separates children of wealthier parents from those of less wealthy parents, and, however unintentionally, divides students by ethnic groups. She argues that to enhance democratization within U.S. schools, education leaders should quit grouping students by skills and organize classes using whole-class instruction.

Arguing on another side are research scholars who believe that grouping students by skill level helps them learn. Kulik (1992), for example, reviews a massive amount of evidence. He concludes that, on average, skill grouping is moderately effective, and especially that "benefits are positive and often large in special classes for the gifted and talented" (p. 41). Kulik stresses the importance of taking full advantage of classes that are grouped by students' skills. He points out that if teachers cover only the standard curriculum, taught to students at all skill levels in the same way, it should not be surprising if students grouped by skill level reap little benefit. Adjusting the curriculum should, according to Kulik, make the grouping more effective.

A third group of scholars argues that although it is important to identify effective ways to group students to enhance learning, traditional grouping practices such as the XYZ method have little effect overall. For example, in an extensive review of traditional grouping practices, Slavin (1993) concludes, "Overall achievement effects were found to be essentially 0 in middle and junior high school grades (6-9). Results were close to 0 for students of all levels of prior performance — high, average, and low" (p. 535). Thus, at least three major lines of argument appear in debates about skill grouping. Lack of resolution of this conflict points to the need for additional empirical evidence.

Our review of the impact of skill grouping focuses mainly on learning outcomes. In addition, several studies assess attitudes and preferences of students, parents, and teachers in between-class or XYZ skill grouping versus whole-class arrangements. We provide results from some of these studies later in this article.

Selection Criteria

We include only studies that provide data from experiments carried out in actual classrooms. How did we choose such investigations from the hundreds of articles, essays, research reports, philosophical and political discussions, and other documents that are now available? We used two criteria: 1) Each study had to be an actual experiment that compares learning in skill-grouped classes with learning from whole-class groupings in a school or several schools — that is, a treatment and a control group; and 2) the study had to be designed as a randomized field trial — the assignment of the treatments (skill grouping versus whole-class grouping) must be either randomized or a close approximation to randomization.

We found several published studies that met these criteria. Several others were unpublished doctoral theses. The studies span a time period of more than fifteen years. Appendix 1 describes our literature search protocol. Computer searches of library databases turned up review articles and original research articles. We also benefited from the advice of colleagues and by hand searching recent journals.

BOX 1
The Importance of Experiments

To learn the consequences of making a change in a complex system, that is, in how we treat something, then it is necessary to actually implement a change in the treatment and measure the effect. We cannot expect reliable results if we only observe different groups. For example, to see if gaining weight will make adults taller, consider measuring heights for people of different weights. Although the result looks as if increased weight might increase stature, many personal experiences with weight gains teach us this is not so. The key point is that we did not implement a change and measure the consequences. We observed people already treated in many ways who were not initially equivalent. Experimentation is one way of making changes and viewing their consequences in a controlled manner.

Let us review what we mean by experimentation.

To test whether one way of doing something is preferable to another, investigators compare the performance of comparable groups treated in the two ways. These groups, called the experimental and control groups, must be equivalent before the treatments are imposed.

In the studies we review in Part 1, the experimental group receives some form of skill grouping and the control group usually receives whole-class instruction. The initial equivalence of the groups is often achieved by randomly drawing the experimental group and the control group from a common pool of students. (Other devices that are equally or almost equally appropriate are sometimes used.)

The notion of an experiment as used here is *not* the common one of tentatively trying out an innovation to see if it will work. Instead, an experiment is a systematic way of carrying out an investigation to find out how well two treatments perform and how much better the winner is. A detailed protocol tells how each step in the investigation should be handled, how the experimental and control groups are formed, and what outcome measures are to be gathered and compared. In such experiments, each treatment that is examined shows some promise from preliminary studies. Investigators want to compare the effectiveness of the treatments.

Our inclusion criteria forced us to set aside many studies. Many studies of skill grouping use no comparison group at all. Other studies employ a "matched" design. Some of these studies compare the performance of students in a school using skill grouping with the performance of students in a seemingly similar school using whole-class instruction. Such matching studies do not guarantee initial equivalence of groups. Randomized field trials generate the strong evidence needed to answer our questions about skill grouping because they assure that the two skill groups initially are statistically equivalent.

Reviews by Kulik (1992) and Slavin (1987, 1990) helped us to identify critical studies of skill grouping and to sharpen our definition of this practice. Their appraisals of the large body of literature provided an initial focus for our work. Kulik also kindly gave us some specialized information.

Comparing Achievement in XYZ Grouping with Whole-Class Instruction

Our literature search turned up only 10 randomized or nearly randomized experiments comparing the effectiveness of XYZ grouping with that of whole-class instruction; all were carried out between 1960 and 1975. We were surprised to find so few randomized investigations and were troubled that the majority of the studies are of modest size and scope. Appendix 2 describes the experimental studies reported in this section.

Each study took place in a single school. In two studies, two grades were involved. Overall, the grades ranged from 3 through 11, and six of the studies included grade 8 or 9. Because only one study used a grade below 7, these studies are primarily associated with middle and senior high schools.

Seven studies focused on a single subject, such as English or mathematics. The other three dealt with a more extensive collection of subjects. In sorting students into skill levels, two or three levels were commonly used, and one study appears to have had as many as nine, though its analysis used only three. Two studies lasted half a year, seven studies lasted a year, and only one lasted two years.

Some investigators initiated studies to learn whether skill grouping could improve the performance of students compared with whole-class instruction. Other investigators intended to demonstrate that little would be lost by giving up skill grouping and switching to whole-class instruction. The total numbers of students involved in the experiments were about 80 students in three studies, about 170 students in two studies, and about 200, 300, 400, 500, and 600 students each in the other five studies, or about 2,600 in all.

For most of these studies, cognitive results could be assessed using an effect-size statistic. The effect size is a positive or negative number that assesses change while taking into account the variability of the performance of the population. In our orientation, positive numbers favor skill grouping, negative numbers favor whole-class instruction, and zero stands for equality. To aid in interpreting effect sizes, we next explain how to translate the effect size into the gain in skill level that a typical student (the median or middle student) in the experimental group would make.

BOX 2
Appreciating Effect Sizes

Although effect sizes of the magnitude of 0.1, 0.2, or 0.3 may not seem impressive gains for a single individual, for a population they can amount to a great deal. A few examples may help.

Example: A 0.20 effect size corresponds in the U.S. to the difference between the average heights of 15-year-old versus 16-year-old girls. For large numbers of girls of each age, this average difference may sound small, but most people notice it.

Example: An effect size of 0.3 corresponds to about 30 points on an SAT verbal or mathematics standardized test.

Example: A 0.80 effect size is widely noticed and would not be missed even by most casual observers of a situation. For example, a 0.80 effect size corresponds to the mean difference between the heights of 13-year-old and 18-year-old girls.

Source: Cohen (1977).

When the learning achievement of children given an experimental treatment advances by a numerical effect size, how much will a typical child — one at the median or 50 percent point of the distribution — move up? If the typical child benefits by an effect size of 0.30, then instead of scoring better than 50 percent of all children, he or she would, according to Table 1, score better than 62 percent of all children. An effect size of 0.10 would move the median child up more modestly, from the 50 percent position to 54 percent.

The tables in Appendix 2 summarize information quantitatively about the design and the findings from the ten studies. For each experiment we provide: author(s), date of publication, grade level of students, class subject, duration of experiment, method of randomization, groupings, number of levels, skill level sample sizes, effect sizes, and non-cognitive findings. The authors of two large studies did not report their findings in a manner that made it easy to summarize their results numerically; therefore, Appendix 2 describes verbally the outcomes of these two studies.

Using data from the research reports, we computed the effect sizes given in Table 2 and in Appendix 2. Table 2 shows both our computed average effect size and the number of students for each of the ten studies. In most instances, the effect sizes are based on the outcomes of standardized tests, though sometimes teacher-made tests were used too.

In Table 3, we summarize the data in a different way. We classify the effect sizes into three groups (positive, near zero, negative) for purposes of simple counts. The choice of ±0.05 as cutoffs for "near zero" effect sizes is arbitrary. The main point is that five of these studies favor skill grouping, three favor whole-class grouping, and two give effect sizes near zero.

Differential Effect of Skill Level

In addition to looking at the overall averages, we also ask how XYZ grouping affects high-skill, medium-skill, and low-skill students. Table 4 gives effect-size estimates for students at each skill level, comparing skill grouping with whole-class instruction. These numbers have to be interpolated in some instances because some studies used only two levels, while others used three or four levels.

Usually, the average of the three effect sizes for the skill levels in Table 4 agrees with the study effect size in Table 2. In the Drews (1963) study, this is not quite so because of the allocation of cases to the levels. In Table 4 we deal with components in the several skill levels, therefore the sample sizes are smaller and the results less stable than the results in Table 2.

When the entries at each skill level are weighted by the sample size, we observe a slight tilt toward skill grouping being more favorable for high-skill than for medium- and low-skill students. The estimates of average effect sizes were 0.08 for high-, -0.04 for medium-, and -0.06 for low-skill groups. These differences are not very reliable, so the observed tilt should be taken as a possibility that skill grouping is slightly favorable for high-skilled students, and slightly unfavorable for medium- and low-skilled students, rather than being taken as a firm research conclusion.

TABLE 1
Effect Size and Percentage Improvement

In reponse to an effect size, the median child (i.e., the child whose performance exceeds that of 50% of the children) improves to exceed the percentage of children shown.

Effect size	.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
Exceeds (%)	50	52	54	56	58	60	62	64	66	67	69
Effect size	.55	.60	.65	.70	.75	.80	.85	1.00	1.50	2.00	
Exceeds (%)	71	73	74	76	77	79	80	84	93	98	

TABLE 2
Average Performance of Skill-Grouped Students as Compared with Whole-Class Grouped Students in the 10 Experiments

Study	*Effect Size**	*No. of Students*
Barton (1964)	.11	204
Bicak (1962)	–.33	75
Drews (1963)	–.04	432
Fick (1962)	.02	162
Ford (1974)	.29	82
Lovell (1960)	.14**	500
Marascuilo & McSweeney (1972)	–.16**	603
Peterson (1966)	–.10	317
Vakos (1969)	.09	184
Wardrop et al. (1967)	.28	82
Sample size weighted average	.00	2641(total)

*Positive effect size favors skill grouping, negative favors whole-class instruction.
** See Appendix 2.

Summarizing the Cognitive Information

Results from the ten studies suggest that XYZ grouping seems modestly preferable to whole-class grouping for high-skill students. In contrast, medium- and low-skill students may learn a little more with whole-class instruction than with skill grouping. Such a conclusion is consistent with the belief that skill grouping benefits only the high-skill students. However, because of variability in the findings of these studies, they do not conclusively favor skill grouping for the high-

TABLE 3
Summary Statistics for Table 2: Number of Studies Comparing XYZ Grouping to Whole-Class Instruction in Three Effect-Size Groups

	No. of Studies
Positive, favoring XYZ grouping	5
Near zero (within 0.05)	2
Negative, favoring whole-class instruction	3

TABLE 4
Effect Sizes for Students at Various Skill Levels

		Effect Sizes[a]	
Study *Skill Level:*	*High*	*Medium*	*Low*
Barton (1964)	.29	.02	.05
Bicak (1962)	–.55	–.33	–.16
Drews (1963)	–.18	.02	–.08
Fick (1962)	.25	.09	–.27
Ford (1974)	.29[b]	.29[b]	.29[b]
Lovell (1960)	.24	.14	.04
Marascuilo & McSweeney (1972)	.03	–.20	–.30
Peterson (1966)	.14	–.42	–.02
Vakos (1969)	.10	.08	.10
Wardrop et al. (1967)	–.01[c]	.42[c]	.42[c]
Sample size weighted average	.08	–.04	–.06

[a] Positive values favor skill grouping and negative values favor whole-class instruction.
[b] High and low not available. Entry is the average for all skill groups.
[c] Pooled medium and low. See Appendix 2.

skill students, nor do they favor whole-class instruction for the other skill levels. These data also indicate an urgent and troubling finding: the effects of XYZ grouping are not very well settled by these investigations. Overall, results of the ten studies suggest that XYZ grouping, on average, does not have much effect on achievement.

For the moment, we note that the studies themselves are small, constrained to single schools for short time periods, and show considerable variability in outcomes. Consequently, we know remarkably little about the impact of XYZ skill grouping on student achievement.

Cross-Grade Grouping: The Joplin Plan

We found only two randomized, controlled studies using the Joplin Plan (see Appendix 3) — a way of grouping students across grades so that they work in small groups with other students who share similar current skill levels. The goal of this plan is to reap the rewards of having students work in a small group of fellow students with similar skill levels, while enabling easy, prompt, upward steps as skills improve.

The two experiments involving the Joplin Plan suggest that it may offer substantial learning benefits. In one experiment on reading skills, Morgan and Stucker (1960) separated 180 fifth- and sixth-grade students into two groups formed from ninety matched pairs. Overall, the Joplin Plan treatment led to reading improvements with an effect size of 0.33. This effect size is larger than those of most of the XYZ skill-grouping studies reviewed earlier.

In a second experiment (Hillson, Jones, Moore, & Van Devender, 1964; Jones, Moore, & Van Devender, 1967), researchers found similarly promising results for the Joplin Plan. The special strength of their study is that it was longitudinal: they followed a group of first-grade children for three years. These children were randomly assigned to two similar groups, starting in first grade.

After eighteen months, students in the Joplin Plan and in the whole-class group were tested in interpreting paragraph meaning, word meaning, and overall reading. The Joplin Plan assignment led to statistically significant differences between the two groups of students. The smallest difference between the two groups was on their comprehension of paragraph meaning, with an effect size of 0.55, favoring the Joplin Plan. For both word meaning and reading, the effect sizes in favor of Joplin were even larger.

After three years, in the follow-up, these differences narrowed. Nonetheless, the findings still favored the Joplin Plan, and with moderately large effect sizes. For example, the smallest difference between the two groups was in paragraph meaning, with an effect size of 0.30. In interpreting word meanings, the Joplin Plan students outperformed the control group by an effect size of 0.38. Again, these effect sizes are larger than those reported in the ten XYZ grouping studies. In the Language Test portion of the Stanford Achievement Test, the effect size at three years was 0.27, favoring the Joplin Plan.

What can we say about the Joplin Plan for assigning students? It is striking that this particular way of grouping students has so rarely been examined using experimental methods. Two randomized, controlled field studies, both with modest sample sizes, do not constitute a substantial commitment to examining the impact of this method of grouping. The results presented here are encouraging, yet our evidence is severely limited.

Within-Class Grouping

We found three experiments that report results on within-class grouping (see Appendix 4). Dewar (1963) compared instruction using within-class skill grouping to whole-class teaching in arithmetic for sixth-grade students. In grade equivalents, the high-, medium-, and low-skill groups each gained about half a grade more than their respective counterparts in whole-class instruction. Dewar's data also reported an effect size of about 0.5; this is a promising finding.

In a second report, Slavin and Karweit (1985) compared a form of within-class skill grouping with whole-class instruction in mathematics. They conducted two experiments, the first with grades 4, 5, and 6, and the second with grades 3, 4, and 5. Each class had two skill groups, the highest 60 percent and the lowest 40 percent of the students. Teachers were trained to push the pace for the high-group and to differentiate materials between the groups. The authors called the skill-grouped teaching Ability Grouped Active Teaching (AGAT), a method developed by Slavin and Karweit (1983) after they examined beneficial features of various teaching methods.

The first experiment had 133 students in the AGAT class and 89 in the control group using the Missouri Mathematics Program (MMP). The design of AGAT is intended to minimize management problems and maintain a high percentage of time on task for the students. The whole-class control groups in both experiments used the MMP, and the second experiment had a further control group that used no special teaching methods. In the second experiment, the AGAT group had 98 students, the MMP had 162, and the whole-class group that received no special teaching methods had 106 students.

The appraised areas of learning in arithmetic were (a) Computation and (b) Concepts and Applications. In the first experiment, when the within-class grouping plan (AGAT) was compared with the whole-class control group (MMP), the effect size was 0.74 for Computation. For Concepts and Applications, the effect size was 0.08. In the second experiment, the effect size for Computation was 0.55. The corresponding effect size for Concepts and Applications was 0.63. In the first experiment, comparing AGAT to a regular whole-class control group yielded an effect size of 0.84. In the second experiment, the corresponding effect size was 0.73.

The gain for Computation is highly statistically significant, but the corresponding gain for Concepts and Applications is not statistically significant. In spite of the lack of statistical significance for Concepts and Applications, the comparative gain is still educationally important.

In a third study, Wallen and Vowles (1960) compared within-class skill grouping to whole-class instruction in sixth-grade arithmetic. They used two schools with two teachers, one male and one female, in each school. They used a special design called a cross-over. Each teacher taught a class one semester using whole-class instruction and the other semester using within-class skill grouping. The sample sizes were twenty-five students per group in School One and thirty-one students per group in School Two.

Two sixth-grade classes were formed in each school by ranking standardized test scores from the previous spring and putting the odd-ranked students in one group and the even-ranked ones in the other. The main finding was that the classes differed in their performance (presumably some teachers were more effective than others), but that the average performance for the two methods of teaching was nearly identical. The classes individually showed little difference when the grouping plan switched from one method to the other.

What does this add up to? Among the three experiments involving within-class skill grouping presented here, two show considerable promise for within-class skill grouping; the other is neutral. We look forward to more extensive studies of this method.

Non-Cognitive Findings from Between-Class Skill-Grouping Research

Our examination of skill grouping focuses on cognitive achievement as measured by standardized tests. We also report on non-cognitive outcomes. Examples of non-cognitive outcomes that we examine include students' evaluation of their own learning; students' perceived anxiety in class; student, teacher, or parental attitudes toward skill grouping; and students' participation (the number of times each student speaks in class) in skill grouping versus whole-class grouping.

Overall, we think that the non-cognitive data tilt in favor of skill grouping. For example, on student self-report measures, skill-grouped students produce higher scores than the whole-class groups, both for liking their school more and for the amount of self-perceived learning. One study examines parents' perception of skill grouping. The findings suggest that parents support having their children skill grouped. Three studies poll teachers on whether they prefer skill grouping. These teachers prefer skill grouping, citing ease of planning the curriculum and classroom dynamics that are more facilitative for learning. The one experiment comparing student behaviors and participation in class finds that the low-skill children who are skill grouped speak up far more and for longer periods than similarly skilled students assigned to whole-class instruction.

Techniques Used for Measurement

Self Reports

In the studies we reviewed, the most frequent technique for measuring non-cognitive effects involved asking a student to respond to a set of statements, such as, "I like the section I am in," and then asking how true that statement is for the student. This method is called self-reporting.

Six experiments reported findings on students' attitudes toward their skill-grouping experience. Students rated how they feel about school, their class placement, and the difficulty of their schoolwork. Four out of six studies found no differences between the skill groups and whole-class groups. The two that found a difference favored skill-grouped students. For example, Marascuilo and McSweeney (1972) found that skill grouping led students to report higher ratings of satisfaction with their class.

Lovell (1960) examined students' attitudes towards school, asking students whether they thought that their teachers enjoyed teaching the class. He found no differences between the skill and whole-class groups in students' positive or negative attitudes toward school, but did find that skill-grouped students believe more strongly that their teachers enjoy teaching. Peterson (1966) found that the low-skill students under whole-group instruction believe that their teachers like them more than do their counterparts under skill-grouped instruction.

Students in skill-grouped classes reported that their classes were more difficult and created more anxiety than their counterparts experienced in whole-group classes. Fick (1962) found that anxiety increased during the year for skill-grouped students, while it decreased for whole-class students, but skill-grouped students rated themselves higher in learning. Ford (1974) found that skill-grouped students perceive their classes to be more difficult than do those in the whole-class group. Bicak (1962) found the opposite.

Teacher Reports

Three researchers — Lovell (1960), Peterson (1966), and Barton (1964) — examined teachers' attitudes toward skill grouping. All three found that teachers prefer skill grouping by a large margin. Teachers repeatedly said that with skill grouping it is easier to plan a curriculum. Teachers also reported that skill grouping creates an incentive for students to push one another to perform. Finally, the researchers found that in any given grade, teachers prefer teaching children who are more highly skilled.

Parent Reports

Barton (1964) examines parents' attitudes toward skill grouping, reporting that 90 percent of parents indicate that they favor having their children in skill-grouped classes. Furthermore, 89 percent of parents do not report anything undesirable about skill grouping. Ninety percent of the parents report that their children had never been teased because they are in a particular skill group; of the 10 percent who say that their children had been teased, most are parents of children in a low-skill group.

Observational Measures

Drews (1963) investigated verbal participation in class by tape-recording and analyzing actual class discussions. She compared the skill groups with whole-class groups on two indices: 1) the number of times each student participated in class discussion, and 2) how long each student talked.

She reported striking results. Skill-grouped students, regardless of whether they were high- or low-skill, spoke more often and longer than students in whole-class groups. Drews demonstrated compellingly for her sample that in whole-class groups, high-skill students dominated the discussion, and that low-skill students tended not to participate actively. In contrast, skill-grouped classes were far more inclusive of students in their discussions (see Table 5). In skill-grouped classes, four times as many of the low-skill students contributed per class, and they spoke twice as long as similar low-skill students in whole-class groups. Further, low-skill students in skill-grouped classes used far more words per contribution (37.4 versus 14.0) than their counterparts in whole-class instruction.

TABLE 5
Number of Contributions per Class

Grouping	*High*	*Low*
Skill	3.44	4.03
Whole-class	7.77	.86

Source: Drews (1963).

Summary of Review for Skill Grouping

After examining fifteen experiments involving skill grouping, we find little evidence that skill grouping has a major impact, either positive or negative, on students' cognitive learning. Further exploration suggests that a few promising methods of teaching may produce substantial effects from certain kinds of skill grouping, but the evidence now available from decades of research is not compelling. This result is dismaying.

Because skill grouping is widely used, the public might reasonably assume that evidence in favor of its effectiveness must be strong. Yet the modest-sized set of investigations just reviewed does not adequately inform educators about the impact, positive or negative, of various forms of skill grouping.

We cannot find a single large-scale, well-designed experiment that follows students over several years to evaluate the impact of skill grouping. In experimentation, short-term studies entail a risk that the existence of the experiment itself may change behavior in a way that leads to a misimpression of the effectiveness of an intervention. The gain observed may disappear after longer experience. In some experiments with skill grouping, the initial effects were larger in the first period of treatment than appeared later. This may have been due to sampling fluctuation, or it may illustrate a problem of the effect of a novel treatment, often called the Hawthorne Effect (see Box 3).

BOX 3
The Hawthorne Effect

Hawthorne was the name of a factory where experiments on the productivity of workers were carried out in the early 1900s. The investigators observed that productivity was influenced by the attention given to the workers as well as by the treatment being investigated, such as changed wages or improved lighting. For example, when the lighting in the factory assembly room was increased, productivity increased, but when the lighting was then reduced, productivity increased again. It is argued that the cause of the original improvement in productivity could not be attributed to beneficial effects of lighting on the production itself, but to the attitudes toward the situation of those participating in the experiment.

The ten randomized, controlled field trials that evaluate the impact of XYZ skill grouping are dated, include few students, and each examines only a single school. The evidence for the effect of XYZ grouping is weak. Only two small randomized studies evaluate the Joplin Plan. Only three evaluate within-class grouping. The evidence about all three forms of grouping is scant.

Our second finding, an average effect size of .00, based on the ten studies of skill grouping using the most widely examined method, XYZ grouping, agrees with Slavin's statement: "The effects of ability grouping were found to be essentially zero for high, average, and low achievers in 27 studies of high methodological quality" (Slavin, 1993, p. 549). It also does not contradict Kulik. He was concerned that if the curriculum did not adjust to skill grouping, and in these experiments it does not seem to, then little benefit from skill grouping could be expected.

Our third finding consists of nuggets of information that are tucked into specific studies. These nuggets may be promising to education policymakers, school leaders, and teachers. For example, although the information available is sparse, the Joplin Plan shows promise, especially for teaching reading. This form of grouping enhances flexibility. Based on the extremely limited evidence now available, it might work especially well for students with less developed skills. Similar remarks apply to within-class skill grouping. But again, the troubling reality is that the extensive research work has not yet been done.

A fourth intriguing observation is that even in a study that finds whole-class instruction to be slightly more effective than skill grouping, students in skill-grouped classes are more engaged with their learning, as measured by how often they speak up in class. This is especially true for the less skilled students. In one study, such students speak nearly five times as often when skill-grouped than their counterparts in whole-class instruction.

A fifth finding shows a slight tilt when examining the impact of XYZ grouping on students with different skill levels. The ten studies give a slight indication that the more skilled students benefit a bit more from skill grouping, while the less skilled students benefit a bit more from whole-class grouping.

We point this out because if this finding holds up under further, careful investigation, leaders of schools may face a dilemma. Well-informed parents of highly skilled children may advocate for schools to skill group their children. Well-informed parents of less skilled children may press for the opposite. This brings us full circle. If these extremely preliminary results do hold up in future larger scale, randomized studies, the Joplin Plan and within-class skill grouping may offer profitable alternatives to traditional XYZ grouping.

Educators in schools that use XYZ grouping may wish to consider Kulik's point about the importance of differentiating the curriculum and the materials in the several skill groups. Such a consideration might lead to a review of whether the presentations and materials are tuned adequately to the skill levels already achieved by the students in the different skill groups.

Among non-cognitive findings, skill-grouped students report better attitudes toward school and perceive that they learned better. Parents are supportive of skill grouping, and teachers prefer it, as reported in the experiments.

To sum up, our main finding does not concern the precise effect sizes for XYZ grouping, the Joplin Plan, within-class, or whole-class grouping. The main finding is that the appropriate, large-scale, multi-site research studies on skill grouping have not yet been carried out, even though the issues have been debated as major public concerns within education for most of this century.

CLASS SIZE

The Tennessee Studies of Class Size

Project STAR (Student/Teacher Achievement Ratio), a study of the educational effects of class size in the state of Tennessee (Word et al., 1994), is one of the great experiments in education in U.S. history. Its importance derives in part from its being a statewide study and in part from its size and duration. But even more important is the care taken in the design and execution of the experiment. Not only are the findings of the experiment valuable, but Project STAR is also extremely important as an example of the kind of experiment needed in appraising school programs, and as proof that such a project can be implemented successfully on a statewide basis.

In a public experiment, it is difficult to stick closely to the protocol of a study because people are bound to have constructive second thoughts after the program begins. For example, in the Tennessee experiment, some changes were made, but cautiously enough not to invalidate the investigation. The main finding was that a small class size in the earliest grades — kindergarten, first, second, and third grades — speeds learning in these years and continues to confer lasting benefits to pupils when they attend larger classes in later grades.

The political atmosphere in Tennessee was favorable to this experiment because then-Governor Lamar Alexander had put education at the top of his agenda for his second term (Alexander was later Secretary of Education in the cabinet of President Bush). The Tennessee legislature and the education community had been motivated by Project Prime Time (Malloy & Gilman, 1989; Tillitski, 1990), a promising study carried out in nearby Indiana examining the benefits of small classes in the early grades. Noting the expense associated with additional classrooms and teachers, the Tennessee legislature decided that it would be wise to have a solid research base before adopting such a major program. At the same time, discussions of the cost and effectiveness of teacher aides in elementary classes adjoined this issue to the class-size investigation. Therefore, the legislature authorized and funded a four-year study of the effects of class size and teacher aides on student learning in the early grades.

The idea that drove the Tennessee study is that in smaller classes, teachers have more time to give to individual children. In addition, teachers and administrators who advocate small classes for students who are beginning school seem to think that they are dealing with a "start-up phenomenon." When children first come to school, they face a great deal of confusion. They need to learn to cooperate with others, to learn how to learn, and to get organized to become students. They arrive from a variety of homes and backgrounds, and many need training in paying attention, carrying out tasks, and engaging in appropriate behavior toward others in a working situation.

In the experimental classes, Tennessee reduced the class size from about 23 to about 15, by approximately one-third, in kindergarten, first, second, and third grades. The children moved into regular-size classes in the fourth grade.

The study was carried out in three kinds of groups: 1) classes one-third smaller than regular-size classes; 2) regular-size classes without a teacher aide; and 3) regular-size classes with a teacher aide. By comparing average pupil performance in the different kinds of classes, the benefits of small classes or the presence of a teacher aide can be assessed.

The experiment, carried out in 79 schools the first year, randomly assigned both children and teachers to the classes; each school had at least one class of each of the three kinds so that comparisons could be carried out within the same school. Otherwise, the effects on the groups of classes might have depended on the properties of the schools presenting the teaching or of the neighborhoods where the children lived. In the second year, the experiment, for example, included 76 schools with 331 classes, including 6,572 children in inner-city, urban, suburban, and rural schools. (The numbers differed a bit from year to year.)

The first phase of Project STAR carried out a four-year statewide experiment with three kinds of classes. After the experiment, a second phase, the Lasting Benefits Study, followed participating children into later grades and recorded their academic progress (Achilles, Nye, Zaharias, & Fulton, 1993; Nye, Zaharias, Fulton, & Achilles, 1993; Nye, Zaharias, Fulton, Achilles, Cain, & Tollett, 1994). A third phase, Project Challenge, initiated in 1989 (Achilles, Nye, & Zaharias, 1995; Nye, Achilles, Zaharias, & Fulton, 1993), implemented the small classes in the seventeen districts with lowest average per capita income among the 139 Tennessee districts.

BOX 4
The Tennessee Class-Size Experiment

The Tennessee project on the effectiveness of small classes and of teacher aides has had, until the present writing in 1996, three phases.

Phase 1. 1985-1989: The education system of Tennessee carried out a four-year *experiment*, called Project STAR, to assess the effectiveness of small classes compared to regular-size classes, and of teacher aides in regular-size classes, on improving cognitive achievement in kindergarten, first, second, and third grades.

Phase 2. 1989-ongoing: The Lasting Benefits Study was an observational study of the consequences of the experimental program on the children when they moved to regular-size classes in the fourth, fifth, sixth, . . . grades. This research phase asked whether the children who started in the smaller classes performed better in later grades. Only students who had been in the experiment (Phase 1) could contribute data to this second phase.

Phase 3. 1989-ongoing: Project Challenge implemented the small classes in kindergarten, first, second, and third grades in the 17 districts of Tennessee where children are highly at risk of dropping out early. The districts have the lowest average incomes in the state.

Major Findings on Class Size

After four years, it was clear from the experiment in Phase 1 that smaller classes did bring substantial improvement in early learning in cognitive subjects such as reading and arithmetic. After following the groups further in Phase 2, the Lasting Benefits Study (Nye et al., 1994), the effects persisted into grades 4, 5, 6, and 7, after pupils moved to regular-size classes, so that students who had been originally enrolled in smaller classes continued to perform better than their grade-mates who had started in larger classes. In the first two years of Phase 1, minority students gained twice as much as the rest, but after that they settled back to about the same gain as the rest. The minority students were almost all African American.

As a consequence of the four-year Phase 1 investigation, the Tennessee legislature decided to implement the small-class program in the seventeen school districts where the children seemed most at risk for falling behind — districts with the lowest per capita incomes. The results of the first three years of this Phase 3 program, called Project Challenge (Achilles, Nye, & Zaharias, 1995), have been encouraging: in the smaller classes, the children from these districts are performing better on both standardized and curriculum-oriented tests than pupils in the same districts in earlier years. Indeed, their end-of-year performance has raised their district ranking in arithmetic and reading from far below the average for all districts to above average.

The presence of teacher aides, though beneficial, did not produce improvements during Phase 1 comparable to the effect of the reduction in class size, nor did their presence seem to have as much lasting benefit during Phase 2.

Discussion and Implications

Of course, after an experiment such as Project STAR reports its results, those hearing of them are likely to say that they already knew what the results would be and therefore that their natural wisdom made this substantial experiment superfluous. In this case, however, we know that the results were not so obvious. Glass and his colleagues (Glass, Cohen, Smith, & Filby, 1982) gathered data on student achievement related to class size and found the literature extremely variable in reported results. By applying a method of research synthesis that they called meta-analysis, they were able to make the case for smaller classes leading to greater achievement. Meta-analysis, however, was not viewed favorably by all professionals then, and the effect of class size continued to be seriously debated. Today, in 1996, meta-analysis is in wide use in medicine and in the social sciences, including education, especially for combining the results of similar randomized controlled experiments (Cooper & Hedges, 1994).

Consequently, the request of the Tennessee legislature for a convincing study should not be regarded as a mere delaying tactic, but as a reasonable request for verification. When the education of children and the use of large amounts of money are at stake, citizens may well ask for assurance stronger than the average citizen's unaided intuition or the specialist's best speculation.

Quantitative Evidence

The Tennessee study was a randomized experiment. What is important about the experiment is that the treatments (small class, regular-size class, regular-size with teacher aide) were randomly assigned by the investigator. Both students and teachers were randomly assigned to the treatment groups. Thus we can be assured that the assignment of treatments did not depend on preferences of teachers, students, or parents. Furthermore, the randomization gave a way of equating the treatment groups before the program began.

Table 6 shows the composition of classes in the schools in the study, broken down by type of location. An important point about the findings of academic gains is that gains from small classes occurred for all types of students in all types of districts.

Although 180 schools offered to participate in the Project STAR, only 100 met the qualifications, and only 79 actually participated in the kindergarten year, the first year of the experiment. The treatments planned for the program started in 1985, beginning with kindergarten and continuing each year through grades 1, 2, and 3. The classes were of three types: 1) small: 13–17 pupils; 2) regular size: 22-25 pupils; and 3) regular size with a teacher aide: 22-25 pupils. The small classes had an average of 15 students, down about 35 percent from the average regular size of 23 students.

TABLE 6
Composition of the First-Grade Cross-Sectional Sample in the Second Year of the Tennessee Experiment

	*Location**			
	Inner-City	*Urban*	*Suburban*	*Rural*
Number of schools	15	8	15	38
Number of classes				
All white students	0	18	28	119
All minority students	65	0	13	0
Mixed classes	5	23	21	39
Total classes	70	41	62	158
Number of students	1495	804	1214	3059

Source: Finn & Achilles (1990). Reproduced, with permission, from their Table 1.

* Legislators did not define the terms inner-city, suburban, urban, and rural schools. The investigators put inner-city and suburban schools in the category of metropolitan areas. Inner-city schools were defined as those in which more than half of the sudents received free or reduced-price lunches. Schools *in the outlying areas of metropolitan cities were called suburbs. In the non-metropolitan areas, schools* in towns of more than 2,500 serving primarily an "urban" population were called urban, and the rest were classified as rural.

In assessing performance, two kinds of tests were used: 1) standardized tests (the Stanford Achievement Test [SAT]), and 2) curriculum-based tests (Tennessee's Basic Skills First Test [BSF]). Standardized tests have the advantage of being used nationwide, but the disadvantage is that the tests are not directly geared to the course of study taught locally. Curriculum-based tests reverse the benefits and disadvantages of standardized tests, measuring more directly the increased knowledge of what was actually taught. Unfortunately, curriculum-based tests usually cannot tell us how the results stand in the national picture.

We can use effect size (recall Box 2), to measure the improvement in performance of one treatment over another. Table 7 shows gains in effect sizes in reading and math for the standardized SAT tests and for the curriculum-based BSF tests in first grade, both for small class versus regular-size class without a teacher aide and for regular-size class with an aide versus regular-size class without an aide. The effect sizes are around 0.25 for small class versus regular-size class without an aide and around 0.10 for regular-size class with an aide compared to regular-size class without an aide. Thus, the small class size advances the typical student an additional 10 percentile points, to the 60th percentile, while the aide advances the same student 4 percent, to the 54th percentile. Although not huge, these improvements are substantial; when applied to a large population, they represent a solid advance in student learning.

One way to summarize these results is to provide the percentiles for the average score based on national norms for the SAT test. Table 8 shows such results for small classes, regular-size classes, and regular-size classes with teacher aide, for both Total Reading and for Total Math. Averaged over the four grades, the small classes gained more than eight percentile points over the regular-size classes without aides in reading and nearly eight percentile points in mathematics. The addition of the aide to the regular-size class results in a slight gain in both reading and math over the regular-size class without the aide.

TABLE 7

Gains in Effect Sizes: 1) from small classes in first grade compared with regular-size classes, both without aides, and 2) from regular-size classes, each with an aide, compared with regular-size classes, each without an aide

	SAT reading	*BSF reading*	*SAT math*	*BSF math*
Small classes compared with regular-size classes, without aides	.30	.25	.32	.15
Regular-size classes with aides compared with regular-size classes without aides	.14	.08	.10	.05

Source: Finn & Achilles (1990). Adapted from Table 5.

TABLE 8
Percentile Based on Stanford's Multilevel Norms

Grade Level	*K*	*1*	*2*	*3*
Total reading SAT (percentile)				
Small	59	64	61	62
Regular without aide	53	53	52	55
Regular with aide	54	58	54	54
Total math SAT (percentile)				
Small	66	59	76	76
Regular without aide	61	48	68	69
Regular with aide	61	51	69	68

Source: Word et al. (1990). Adapted from data given in their Figures 1 and 2.

An encouraging finding is that students' early experience with the smaller class size has had a lasting effect beyond the moment when the children moved to regular-size classes. In a paper presented at a meeting of the North Carolina Association for Research in Education at Greensboro, North Carolina, Achilles, Nye, Zaharias, and Fulton (1993) reported on the three-year follow-up study (Lasting Benefits Study) of the Project STAR experiment. These authors found that in the fourth and fifth grades, the children who had originally been in small classes scored higher than those who had been in the regular-size classes or the regular-size classes with an aide. In the fourth grade — the first year after moving to regular-size classes — the effect size was about 0.12 averaged across six different cognitive subjects studied, and in the fifth grade, the effect size was nearly 0.20.

In the seventeen Project Challenge districts implementing small classes in Phase 3, both the reading scores and the math scores improved over the next three years, compared to previous performance of children in these districts. The gains in effect sizes were 0.4 for reading and 0.6 for mathematics. Before the small classes were introduced, these districts had been performing well below the average for the state in mathematics; after the intervention, their performance moved above the average. It should be understood that the gains recorded here are not part of an experiment; they are consequences of implementing the program. The comparisons, then, are not as well equated as they were in the original investigation. To make sure the gain was a result of the smaller classes, we would have to carry out a new experiment in the districts where the plan was implemented.

An additional way to report the progress of students in the districts in Phase 3 is to provide the average rank of the test scores in reading and mathematics for the seventeen Tennessee districts in Project Challenge for the years reported

TABLE 9
Grade 2 Average Ranks of Test Scores for the 17 Districts among the 138 School Districts for Early Years of Project Challenge

	1989–1990	*1990–1991*	*1991–1992*	*1992–1993*
Reading average rank	99	94	87	78
Mathematics average rank	85	79	60	56

Source: Achilles, Nye, & Zaharias (1995, Appendix B).

so far (1989–1993). The scores are a mixture of both the SAT and the BSF tests. Achilles, Nye, and Zaharias (1995) report the results shown in Table 9 for second-grade students. (The total number of districts changed to 138 rather than 139.) The average rank for all districts is 69 (midway between 1 and one 138); note that small ranks mean better scores (i.e., nearer the top of the rankings). In mathematics the average rank for 1991–1992 and for 1992–1993 is below 69 (consequently the district scores are above the median rank) so that the seventeen districts show the startling improvement of a gain of 21 ranks in reading and 29 in mathematics for grade two over a three-year period. The same report mentions that the corresponding grade-one analysis shows that the seventeen districts rank better than the state average in *both* reading and mathematics in 1992 (see Appendix B footnote in Achilles et al., 1995).

In total, the evidence is strong that smaller class size at the beginning of a child's school experience does improve performance on cognitive tests. The Lasting Benefits Study confirms that the effect continues into later years when children are placed in regular-size classes. In addition, the implementation of the program for the economically poorer districts seems to be improving their children's performance by noticeable amounts.

(A more detailed non-technical report of Project STAR mentioning some special difficulties is available in Mosteller, 1995).

Other Issues

Policy Is Not Automatic

When a well-designed and implemented study comes out with a definite finding, people sometimes believe that the finding should have automatic consequences for policy. Of course, that is not true. The policymaker has to give serious consideration to all the available alternatives, and to the costs and social consequences of implementing the new policy suggested by the findings. For example, other interventions may work better than the one being presented. If so, are they cost-effective? Even if the treatment is valuable, one may ask whether it is something to be given to every person or even to any person. (Some medical treatments are so expensive no one can afford them, for example.) The class-size study in Tennessee is a good example, both theoretically and in practice, of how a policy decision may be made based on the definitive results of a well-designed and implemented experiment.

For example, after finding out that smaller class size worked, Tennessee policymakers might have considered reducing class size in all the classes in kindergarten through grade 12. Such action was not in the spirit of the investigation, however, which was intended to find out whether early treatment would improve the performance of children, not only during the initial years, but also after they moved to regular classes. Thus, in the case of the Tennessee treatment, if policymakers decided to implement the smaller class in grades K-3 in every school in the state, that would mean a class size reduction in about 30 percent of all the classes, K-12. But instead, policymakers asked themselves where it would likely be most effective to introduce this intervention and decided to implement it in the seventeen districts with the lowest per-capita income. Thus they decided to use the method in about 12 percent of the state's districts. All told, then, they reduced class size in only about 4 percent of all K-12 classes statewide; there was no leap to use smaller classes in every classroom in the state, nor even in all districts, for the first four grades (K-3). By targeting and restricting use of an intervention, society may find its partial use affordable.

It is important to monitor the outcome of this intervention for the children beyond the first four years to see whether there is lasting benefit for the group being specially treated now, namely those students in the state's poorest districts. One can imagine that the effect might wear off after a while. The opportunity to study the effectiveness of the intervention in a group especially needing it should not be missed.

What Is the Optimum Size Class?

The question of optimum class size is an open one, and we do not have information from this investigation on a variety of sizes of classes. Within the range of what is affordable, we now have reason to believe that smaller classes are preferable for young students in grades K-3. But *some* desired training probably cannot be accomplished in classes as small as one or two students, even if such classes were affordable. Learning to work in a group, for example, as students must in school, requires participating in a group.

Summary of Review for Class Size

The most important aspects of the Tennessee studies on class size flow from the fact that a large, sustained, randomized, controlled experiment was carried out, and that it provided substantial and definitive findings. Such investigations give other educators something to emulate. As we discuss in Part 3, such emulation is much needed.

Much of the strength of the Tennessee study comes not from size alone, but also from its inclusion of a variety of schools, from the different mixtures of students in these schools, and from its statewide nature. Additionally, the study continued over several years. This work has taught us that there seems to be a definite effect of class size.

What is so special about the Tennessee studies? Not only did they carry out a large experiment, but they followed up to see what happened to students originally taught in smaller classrooms and found a persistent favorable effect. The state then actually implemented a change — they introduced small classes in the state's seventeen lowest income districts. In their follow-up, they found that end-of-year grades for these districts improved, so much so that their end-of-year scores rose above the mean rank of scores for the state in reading and mathematics for these early-grade students. Thus, three lines of evidence persuade us that small classes improve learning and that this improvement persists in later grades, when students move to classes of regular size. (At this writing, we have no data from the seventeen low-income districts after their students moved to regular-sized classes.)

Although the finding that class size matters is important, and the size of its impact makes class size an attractive variable to adjust, we want to stress a different point. Had Tennessee found the opposite — that smaller class size did *not* improve performance — that finding, while disappointing, still would have been valuable. What is most important is that the study was carried out with regard to design, numbers, and variety of students and teachers so that the results are firm — that is, generalizable to other schools in other states. The study surely has more general applicability than to Tennessee alone.

The findings do not automatically mean that reducing class size is the best way to improve schooling. They do suggest, however, that to be accepted, an equally expensive proposed innovation should have strong evidence of being as effective as a reduction in class size.

LESSONS FROM THE TWO REVIEWS

We present these two reviews — of skill grouping and of class size — for two reasons. First, each review is important in its own right, posing a complex policy question for practitioners. Educators work with scarce resources and constrained budgets, and must decide how to organize students into classrooms. Deciding on how to group diverse students should be done with care and thought. Having access to strong research and policy studies will enable educators to make wise choices.

The second reason for presenting these reviews is to illustrate in a dramatic way a great challenge for our nation's education system. For educators to make wise choices, they must be confident that such choices are based on sound evidence. Hunches, anecdotes, and impressions may have been the only available options in the year 1900, but as we approach the year 2000, society has a broad set of analytic design techniques, widely accepted and effectively used in many fields, that can offer more reliable evidence than hunches and impressions. These two reviews demonstrate the large gap in knowledge that can emerge between answers to a question that is investigated in a substantial way, as class-size illustrates, and a question that is investigated unsystematically, as skill-grouping illustrates.

Educators develop many original ideas for improving schools, and some of them are implemented. What has not ordinarily been done is to study innovations in education in a sustained way, both to improve a new idea and to provide evidence that it is more effective than other approaches with the same aim.

Although small-scale studies are done, and frequently done very well, the field of education initiates few large-scale studies that are controlled experiments or close substitutes for them. The few large studies now available have mainly been sample surveys or observational studies, such as the National Assessment of Educational Progress (Jones, 1996). These studies primarily assess the state of student performance, rather than compare methods of teaching or organizing students.

Not all questions can be tackled using controlled experiments, but many can be. We need larger scale investigations because studies carried out in single schools always have the limitation of doubtful generalization. Studies carried out in a single semester or a single year suffer from a similar weakness. The Tennessee class-size studies were carried out in many schools with classes of differing composition over a period of several years. And the information came 1) from the experiment itself, 2) from the follow-up study of the experimental students, and 3) from experience after its program was put into practice. Thus, size of investigation, diversity of schools and students, duration of the investigation, and variety of sources of information, as well as the critical feature of randomization, all contribute to our appreciation of what happened when the size of classes was changed.

Although one could take the view that the political sensitivity of skill grouping makes it difficult to carry out extensive research, that position would miss the point of our discussion. The general point is that large enterprises need to evaluate their activities systematically, and to review potential new interventions on a regular basis, both to improve them and to compare their effectiveness with that of other innovations. To be effective, the evaluations need to be large enough to come to definite conclusions about the merits of an intervention.

These reviews illustrate just two examples of dilemmas that have been familiar to educators for over a century. Many other important questions also could use sustained inquiry. They include: "What is an appropriate amount of homework in different classes for children at different ages?" "How should we distribute time on task among different school subjects?" "Will adding 50 percent to the hours spent on a school topic lead to a comparable gain in learning, and, if so, in what sense? Better retention? More ground covered? Improved ability to use the material in practice?" "During summer months and vacations, are students losing too much of what has been learned in the school year?" "How can we best address issues of civility, safety, and violence in the schools?" "When children start school without knowing the English language, what process of language instruction can both maintain progress in school subjects and still lead to fluent

English performance?" The shortage of compelling answers to such questions illustrates how educational practice needs to benefit from a more extensive evidential base.

The main contribution of our examination of the literature on skill grouping is a sharpened awareness of the limited amount of rigorous investigation that has been done. Evidence for or against particular approaches does not overwhelm us. Because every school must deal with the distribution of students among classes, one would have expected extensive and sustained experimentation. As populations change and technologies for teaching change, new rounds of experimentation for each generation should be conducted to evaluate the effectiveness of teaching in different ways under new circumstances.

Currently the United States has no systematic national program for evaluating the impact of different education policies or teaching practices. By contrast, carefully designed experiments are carried out routinely in the evaluation of therapeutic drugs, because such investigations are required of pharmaceutical companies by the Food and Drug Administration to have their products approved for marketing. Clearly, the United States needs analogous programs of research on the effectiveness of educational interventions. An industry that serves 44 million students and employs 2.5 million teachers at a cost of over $300 billion each year needs programs for research on the effectiveness of its methods, not just information about numbers of students and their annual performance. Not only are we lacking the strong information that is needed — but also no process is yet in place for acquiring it.

We learned from our review of XYZ grouping that we cannot find a single large-scale, well-designed investigation that follows students over several years to evaluate its impact. Indeed, we cannot find a single large-scale investigation that follows a variety of students over a *single* year. And while potentially exciting new ways of grouping students are now being suggested and actually implemented in some schools, much more information is needed about these innovations to reach firm conclusions about their effectiveness and their generalizability.

What is so critical about large-scale studies? To work in many parts of our country, any innovation must be adaptable to different populations of children, teachers, and parents. Occasionally we have the opportunity to delight in news of breakthroughs by charismatic teachers and education leaders. When they occur, we should applaud and support the results and the innovations. But, charismatic leadership is hard to export, and the leaders often must move on to other good works. When they leave, their programs are rarely able to maintain such a high level of performance.

It is important that the educational policies we evaluate be strong enough to maintain a consistently high level of performance when the initiators move on to other tasks. In short, we are looking for educational interventions that work for varieties of populations of students, teachers, and parents. These interventions might be characterized as "robust." We are not arguing that one organizational plan must work for all schools, but rather that in a national system we need a few methods that rest on thorough evaluation, and that seem to work well in a variety of circumstances.

The examination of XYZ grouping illustrates this lack of thorough evaluation. For example, half of the studies we report are doctoral dissertations by students in the field of education. Such studies are valuable, and some are path-breaking. But a national education system should not expect the hard work of a handful of largely unfunded doctoral students, together with their professors and friends in nearby school systems, to be the equivalent of, or a substitute for, a national program of research in education.

Why don't substantial investigations occur more frequently? One tempting explanation used by critics is to blame the "education establishment." But it is no help to blame any subgroup of practitioners or research specialists. We wonder, rather, about our nation's commitment. Perhaps a brief look at how our federal government allocates funds and supports research and development will help to clarify the situation. If we ask what fraction of total federal expenditures in several fields are specifically earmarked to support research and development, we find that for health, the research and development portion of its budget is more than 13 percent. For defense it is more than 12 percent. For space exploration it is about 50 percent. For energy utilization it is about 55 percent. And for education it is less than 1 percent (Berryman, 1995). Education is in a class by itself. Unenviably.

One may ask whether large studies existed many years ago. Yes, we had large studies. However, they were not randomized field experiments, but sample surveys. Surveys answer questions about what happens to different groups of students who attend different sorts of schools. Inferences about causality are rarely compelling when they come from surveys. Randomized, controlled field trials are needed instead. Understanding the importance of trying different treatments on comparable groups to establish their effects has migrated as a critical idea from agriculture to medicine. It needs to push forward more strongly into education.

We need more investigations of the kind carried out in Tennessee, where school districts across a state cooperate to contribute to an important finding. One can envision collections of districts or states joining together to design studies of mutual interest, just as medical institutions now routinely join together to carry out cooperative randomized clinical trials. The medical and health care communities have come to expect this. The education community should expect no less. The National Academy of Education discusses extensively the need for and value of research in their report entitled *Research and the Renewal of Education* (National Academy, 1991), and they propose a national research agenda in five educational areas.

Our hope for policy research in education is that leaders at state and national levels, as well as practitioners and academics, will increasingly appreciate the importance of basing policy decisions on evidence from large-scale, sustained, carefully designed studies. To a cynic who says "It is all too complicated," our response is that if systematic, long-term field trials can be initiated in health and medicine and welfare reform and job training, they can be initiated in education as well. We look forward to a time when several states or a group of districts or national organizations initiate a well-designed substantial policy analysis of how to organize students among classes to enhance their learning.

A recent report on productivity in education (Berryman, 1995) points out that real spending per student on education increased by 31 percent from 1975 through 1991. Yet by most measures of performance, U.S. students are not improving their achievement nearly as much, though they do not seem to be losing ground either. A good way to improve performance is to initiate a long-term, sustained program of policy analyses. This may be the best way to help educators who, out in the field in their schools, must implement on a daily basis decisions about such matters as class size, class organization, amount and type of homework, curriculum, and how to integrate modern technology. These educators need a solid source of evidence to help them in their decisionmaking. Supplying compelling evidence-based information to educators about teaching practices should yield benefits that make a real difference.

References

Achilles, C. M., Nye, B. A., & Zaharias, J. B. (1995, April). *Policy use of research results: Tennessee's Project Challenge.* Paper presented at the Annual Convention of the American Educational Research Association, San Francisco.

Achilles, C. M., Nye, B. A., Zaharias, J. B., & Fulton, B. D. (1993, January). *The Lasting Benefits Study (LBS) in grades 4 and 5 (1990–1991): A legacy from Tennessee's four-year (K-3) class-size study (1985–1989), Project STAR.* Paper presented at a meeting of the North Carolina Association for Research in Education (NCARE), Greensboro, NC.

Barton, D. P. (1964). *An evaluation of ability grouping in ninth grade English.* Unpublished doctoral dissertation, Brigham Young University.

Berryman, S., with others (1995). *Using what we have to get the schools we need: A productivity focus for American schools.* New York: Columbia University, Teachers College, Institute of Education and the Economy, Consortium on Productivity in the Schools.

Bicak, L. J. (1962). *Achievement in eighth grade science by heterogeneous and homogeneous classes.* Unpublished doctoral dissertation, University of Minnesota.

Chubb, J., & Moe, T. M. (1992, September). *Politics, markets, and equality in schools.* Paper delivered at Annual Meeting of the American Political Science Association, Chicago.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Dewar, J. A. (1963). Grouping for arithmetic instruction in sixth grade. *Elementary School Journal, 63,* 266–269.

Drews, E. M. (1963). *Student abilities, grouping patterns, and classroom interaction* (Final Report of Cooperative Research Project No. 608, The Effectiveness of Homogeneous and Heterogeneous Ability Grouping in Ninth Grade English Classes with Slow, Average, and Superior Students). East Lansing: Michigan State University, Office of Research and Publications.

Fick, W. W. (1962). *The effectiveness of ability grouping in seventh grade core classes.* Unpublished doctoral dissertation, University of Kansas.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557–577.

Ford, S. (1974). *Grouping in mathematics: Effects on achievement and learning environment.* Unpublished doctoral dissertation, Yeshiva University.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size: Research and policy.* Beverly Hills, CA: Sage.

Hillson, M., Jones, J. C., Moore, J. W., & Van Devender, F. (1964). A controlled experiment evaluating the effects of a non-graded organization on pupil achievement. *Journal of Educational Research, 57,* 548–550.

Jones, J. C., Moore, J. W., & Van Devender, F. (1967). A comparison of pupil achievement after one and one-half and three years in a nongraded program. *Journal of Educational Research, 61,* 75–77.

Jones, L. V. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25*(7), 1–8.

Kulik, J. A. (1992). *An analysis of the research on ability grouping: Historical and contemporary perspectives* (Ability Grouping Research-Based Decision Making Series, No. 9204). Ann Arbor: University of Michigan.

Lovell, J. T. (1960, March). The Bay High School experiment. *Educational Leadership, 17,* 383–387.

Malloy, L., & Gilman, D. (1989). The cumulative effects on basic skills achievement of Indiana's Prime Time: A state sponsored program of reduced class size. *Contemporary Education, 60,* 169–172.

Marascuilo, L. A., & McSweeney, M. (1972, January). Tracking and minority student attitudes and performance. *Urban Education, 6,* 303–319.

Morgan, E. F., Jr., & Stucker, G. R. (1960). The Joplin Plan of reading vs. a traditional method. *Journal of Educational Psychology, 51,* 69–73.

Mosteller, F. (1995). The Tennessee Study of Class Size in the early school grades. *The Future of Children, 5*(2), 113–127.

National Academy of Education. (1991). *Research and the renewal of education: A report, project on funding priorities for educational research.* Palo Alto, CA: Stanford University, School of Education.

Nye, B. A., Achilles, C. M., Zaharias, J. B., & Fulton, B. D. (1993). *Project Challenge, third-year summary report: An initial evaluation of the Tennessee Department of Education "at risk" student/teacher ratio reduction project in seventeen counties, 1989–90 through 1991–92.* Nashville: Tennessee State University, Center for Research in Basic Skills.

Nye, B. A., Zaharias, J. B., Fulton, B. D., & Achilles, C. M. (1993). *The lasting benefits study: A continuing analysis of the effect of small class size in kindergarten through third grade on student achievement test scores in subsequent grade levels. Sixth grade* (Technical Report). Nashville: Tennessee State University, Center for Research in Basic Skills.

Nye, B. A., Zaharias, J. B., Fulton, B. D., Achilles, C. M., Cain, V. A., & Tollett, D. A. (1994). *The lasting benefits study: A continuing analysis of the effect of small class size in kindergarten through third grade on student achievement test scores in subsequent grade levels. Seventh grade* (Technical Report). Nashville: Tennessee State University, Center of Excellence for Research in Basic Skills.

Oakes, J. (1986). *Keeping track: How schools structure inequality.* New Haven: Yale University Press.

Peterson, R. L. (1966). *An experimental study of effects of ability grouping in grades seven and eight.* Unpublished doctoral dissertation, University of Minnesota.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57,* 293–336.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60,* 471–499.

Slavin, R. E. (1993). Ability grouping in the middle grades: Achievement effects and alternatives. *Elementary School Journal, 93,* 535–552.

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston: Allyn & Bacon.

Slavin, R. E., & Karweit, N. L. (1983). *Ability grouped active teaching (AGAT): Teacher's manual.* Baltimore: John Hopkins University, Center for Social Organization of Schools.

Slavin, R. E., & Karweit, N. L. (1985). Effects of whole class, ability grouped, and individualized instruction on mathematics achievement. *American Educational Research Journal, 22,* 351–357.

Slavin, R. E., Madden, N. A., & Leavey, M. (1984). Effects of team assisted individualization on the mathematics achievement of academically handicapped and nonhandicapped students. *Journal of Educational Psychology, 76,* 813–819.

Tillitski, C. (1990). The longitudinal effect size of Prime Time, Indiana's state sponsored reduced class size program. *Contemporary Education, 62,* 24–27.

Vakos, H. N. (1969). *The effect of part-time grouping on achievement in social studies.* Unpublished doctoral dissertation, University of Minnesota.

Wallen, N. E., & Vowles, R. O. (1960). The effect of intraclass ability grouping on arithmetic achievement in the sixth grade. *Journal of Educational Psychology, 51,* 159–163.

Wardrop, J. L., Cook, D. M., Quilling, M., & Klausmeier, H. J. (1967, November). *Research and development activities in r & i units of two elementary schools of Manitowoc, Wisconsin, 1966–1967* (Technical Report No. 35). Madison: University of Wisconsin, Wisconsin Research and Development Center for Cognitive Learning.

Word, E., Johnston, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Lintz, M. N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/Teacher Achievement Ratio (STAR), Tennessee's K-3 class size study: Final summary report, 1985–1990.* Nashville: Tennessee State Department of Education.

Word, E., Johnson, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Achilles, C. M., Lintz, M. N., Folger, J., & Breda, C. (1994). *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) project: Technical report 1985–1990.* Nashville: Tennessee State Department of Education.

This research was supported in part by a grant from the Andrew W. Mellon Foundation to the American Academy of Arts and Sciences project *Initiatives for Children* for its Center for Evaluation.

James Kulik and William Shadish kindly gave us advice and materials. Cleo Youtz and Marjorie Olson handled the manuscript preparation and aided with the references.

John. D. Emerson, Nathan Keyfitz, Martin McIntosh, Edward Miech, Bill Nave, Marjorie Olson, and Cleo Youtz generously provided comments on drafts of the manuscript.

APPENDIX 1
Literature Search to Find Experimental Research on Skill Grouping

The debate surrounding the merits of skill grouping has continued for over one hundred years. In his quantitative review of research on ability grouping, Kulik (1992) identified over 143 studies on the topic dating back as far as 1893. Accompanying the multiple research studies were attempts by scholars to summarize the "current knowledge" on ability grouping. These reviewers carefully weighed the evidence on skill grouping and drew conclusions about its overall effectiveness. Three of the most recent grand summaries, Kulik (1992) and Slavin (1987, 1990), were particularly helpful in identifying the studies that we selected for our review. These authors conducted quantitative reviews and meta-analyses on what they considered to be the most methodologically sound studies on ability grouping. Each conducted a special subanalysis that focused specifically on studies that were experimental, in which the students were randomly assigned to either a treatment or control group. In his review of XYZ grouping, Kulik (1994) identified ten experimental studies. In his review of the effectiveness of ability grouping for elementary students, Slavin (1987) identified five experimental studies. In his later review of skill grouping and its impact on secondary student achievement, Slavin (1990) identified six experimental studies.

We examined all of the studies that were published in journals or on ERIC microfilms. Moreover, several dissertations were kindly given to us by the psychologist William Shadish. In addition to the summary articles, we reviewed over sixty dissertation abstracts from 1920 to 1994 and obtained the several studies that appeared to be experimental. This effort provided no additional experimental studies that used random assignment.

To hunt for studies that might have been published prior to or after Kulik's 1992 meta-analysis, we conducted a further literature search. We used three sources to identify material. First, we conducted a computerized search on the ERIC system, using the key words "Ability Grouping," "Homogeneous Grouping," "Tracking," and "Curriculum Differentiation." We cross-indexed these words with research, experiment, and random assignment from the years 1966 to 1994. Moreover, we examined six major education journals — *American Educational Research Journal, Educational Evaluation and Policy Analysis, Journal of Educational Measurement, Review of Educational Research, School Review,* and *Teachers College Record* — from 1993 to 1995 to identify new studies that might not have been updated into the ERIC system. We did not identify any additional experimental studies that met our selection criteria as a result of this extended literature search.

APPENDIX 2
Ten Experimental Studies Comparing Student Performance under XYZ Grouping with Performance under Whole-Class Instruction

1. BARTON, D. P.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1964	9	English	1 year

Randomization: Teachers used eighth-grade performance, classroom tests, observation, student cumulative record, and consultation with prior teachers to rank students from 1 to 229. Odd ranks were assigned to skill groups, even to whole-class groups. The students were divided by rank into quartiles. The whole-class group was divided into four subgroups so as to maintain a balance of skill levels in the four classes.

		Skill-level sample sizes				
		Quartiles				
Grouping	*No. of levels*	*1*	*2*	*3*	*4*	*Total*
Skill	4	26[a]	26[b]	25[c]	24[d]	101
Whole-class	4	28[a]	25[b]	27[c]	23[d]	103
Effect size[e]		.32	.12	−.07	.08	.11

[a] highest ranked ¼ of students

[b] next ¼ ranked students—second quartile

[c] next ¼ ranked students—third quartile

[d] lowest ¼ ranked students—fourth quartile

[e] Positive is favorable to skill grouping, negative to whole-class instruction.

Non-cognitive findings: 90% of parents preferred children in a "like" ability group. Teachers (who taught both kinds of classes) preferred skill-grouped classes.

* * *

2. BICAK, L. J.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1962	8	Science	2 quarters

Randomization: Each student was randomly assigned to one of three sections. One section was assigned to whole-class instruction, the other two were separated into High and Low based on the Lorge-Thorndike Intelligence Test, Level 4, the verbal form. High skill was defined as scoring above the median of all the students. (Some analyses broke both highs and lows into two groups, but these splits are not used here.)

Grouping	*No. of levels*	*Skill-level sample sizes*		*Total*
		High	*Low*	
Skill	2	23	25	48
Whole-class	2	13	14	27
Effect size*		–.50	–.16	–.33 overall

* Positive is favorable to skill grouping, negative to whole-class instruction.

Non-cognitive findings: Whole-class students stated that they had to spend more time on their science class to the neglect of other topics. In addition, the low-skill students in both the skill-grouped and whole-class instruction did not like their sections when compared to the ratings of the medium-skill and high-skill groups.

* * *

3. DREWS, E. M.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1963	8	English	1 year

Randomization: Using a variety of inputs, students were categorized into three skill levels: High, Medium, and Low. To make whole-class groups similar to the usual composition with 5 to 6 students High, 20 to 25 Medium, 4 to 5 Low, students were drawn randomly from the pool to create these approximate numbers (exactly how is not described, but several methods are available). Eight classes for whole-class instruction were formed. Then using the skill-level stratification, the remaining students were skill grouped into four High classes, six Medium classes, and four Low classes. (The ratios given for "usual composition" do not agree with the composition in hand after the reassignment using many inputs. The randomization procedure balances for statistical comparisons between the groups, but resulted in percentage compositions of High, Medium, and Low that differ in the two kinds of grouping.)

Grouping	*No. of levels*	*Skill-level sample sizes*			*Total*
		High	*Med.*	*Low*	
Skill	3	78	114	59	251
Whole-class	3	23	137	21	181
*Effect size**					*weighted overall***
Language		–.35	.25	–.15	.04
Reading		0	–.20	0	–.12
Average		–.18	.02	–.08	–.04

* Positive is favorable to skill grouping, negative to whole-class instruction.

** Weights proportional to total numbers in the skill levels.

Non-cognitive findings: In the whole-class groups, the low-skill students participated in class discussions much less often than the more skilled, while in the skill-grouped classes, the participation rates of the various skill levels was nearly equal. Low-skill students in skill-grouped classes rated themselves higher as school learners than did the corresponding whole-class students. The three skill-grouped levels rated themselves as nearly equal in ability, while those in the whole-class groups saw themselves as differing substantially, the low-skill giving themselves low ability rating and the more skilled giving themselves higher ratings.

* * *

4. FICK, W. W.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1962	7	Core	1 year

Randomization: Students were ranked according to their scores on the California Short Form Test of Mental Maturity, with order of tied scores randomized with random numbers. By counting down the ranks by thirds, three skill levels (High, Medium, Low) were formed. Using random numbers, each skill level assigned half its sample to whole-class grouping and half to skill grouping.

		Skill-level sample sizes			
	No. of	*Thirds (approximate)*			
Grouping	*levels*	*High*	*Med*	*Low*	
Skill	3	27	27	27	
Whole-class	3	27	27	27	
Effect size*		.25	.09	−.27	.02 overall

* Positive is favorable to skill grouping, negative to whole-class instruction.

Non-cognitive findings: Test anxiety was higher in the skill-grouped classes than in whole-class instruction. However, skill-grouped students had higher ratings on self-perceived learning.

* * *

5. FORD, S.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1974	9	Math	1 year

Randomization: After taking the mathematics part of the Differential Aptitude Test, the population was split into High and Low. Then, one-third of each group was randomly assigned to whole-class instruction, the rest to High and Low skill-grouped instruction. (An extra dimension of compatibility was also used in the design, but is not treated here.)

	No. of	*Skill-level sample sizes*		
Grouping	*levels*	*High*	*Low*	*Whole-class*
Skill	2	22	30	
Whole-class	2	?	?	30
Effect size*		**	**	overall .29**

* Positive is favorable to skill grouping, negative to whole-class instruction.

** The effect sizes for levels could not be obtained because needed information is not available.

Non-cognitive findings: Students in whole-class instruction perceived more class friction than skill-grouped students. Both kinds of low-skill groups perceived the pace of their classes as the slowest compared to the other classes.

* * *

6. LOVELL, J. T.

Date of publication	*Grade of students*	*Class subjects*	*Duration of experiment*
1960	10	Algebra, Biology, English	1 year

Randomization: Sophomore class was ranked by ability (possibly in each subject separately — algebra, biology, and English — but the report does not tell). Even-numbered students were assigned to control classes, odd-numbered to experimental classes, 250 in each type. In the experimental group, subgroups of thirty were formed successively, starting with the highest scores. The nine groups were analyzed according to thirds: High, Medium, Low. In the control groups, students were placed to maximize the variability within the groups. They contained "a balance of exceptional, average, and below average students" (p. 383). Each teacher taught classes in both treatment groups.

Grouping	*No. of levels***	*Sample sizes*
Skill		250
Whole-class		250

** 9 for classes, 3 for analysis

In English, the effect size was 0.25 (also interpretable as the additional fraction of a school year gained by the skill-grouped student over those assigned to whole-class instruction). In both algebra and biology the effect was favorable to skill grouping, but inadequate data were given to report the magnitude.

The skill-grouped students scored higher than the corresponding whole-class students. And among the skill-grouped students, the high-skill ones excelled over their counterparts in whole-class instruction more than the low-skill students excelled over theirs. Based on these quantitative and qualitative findings, we assigned the overall effect size of .14 to Lovell.

Non-cognitive findings: The skill-grouped students gave a higher rating to their teacher's interest in teaching English than did whole-class groups. Similarly in biology, the skill-grouped students gave their teachers higher ratings for interest in teaching. Both at the beginning and the end of the year the teachers preferred skill grouping.

* * *

7. MARASCUILO, L., & McSWEENEY, M.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1972	8, 9	Social Studies	2 years

Randomization: On the basis of several measures of ability and achievement, students were placed in three groups relevant to this study: High (college preparatory), Medium (also college preparatory), Low (not college preparatory), consisting of 35%, 40%, and 20% of the school population. The other 5% are not in the study, nor were certain gifted students. Among those whose parents volunteered to have their children in the study, students were randomly assigned so that the whole-class groups (4 of size 28) had the proportions of the three skill levels in the entire eighth grade. The remaining volunteers formed skill groups: 6 High classes of 32, 7 Medium classes of 32, and 3 Low classes of 25. The whole-class groups had 10 High students, 11 Medium, and 7 Low.

	No. of	*No. of classes and their sizes*		
Grouping	*levels*	*High*	*Medium*	*Low*
Skill	3	6 @ 32	7 @ 32	3 @ 25
Whole-class	3		4 @ 28	

Because the school regularly employed skill grouping, parents had to volunteer permission for their children to be placed in whole-class groups. This policy leads to problems of reporting because comparisons should be made among equivalent volunteering groups.

Two tests were administered. For low-skill students, the whole-class volunteers scored significantly higher on the teacher-made U.S. Constitution Test than the low-skill-grouped students.

The summary says that the paper did not try to find out which method of grouping resulted in higher academic achievement, but whether skill grouping was necessary for effective instruction in eighth- and ninth-grade social studies classes in Berkeley, California. The authors say that the answer was no. "Heterogeneous grouping in a single course had at least a neutral effect, and at best a positive effect, on the cognitive performance of the volunteer students" (p. 318).

Because the instances of volunteers being compared to volunteers are not systematically identified, and because the corresponding numerical findings are not reported, we do not have reliable summary figures.

Teachers had said that they were not well prepared for whole-class instruction in the first year and that they had more time to prepare in the second year.

Two achievement tests were used:

Cooperative Social Studies Test (a standardized test)

Year 1: The investigators broke the items into those taught and those not taught in the curriculum. In both parts of the test, among the skilled groups the high-skill students scored higher than the high-skill students in whole-class instruction, but not significantly higher at the 5% level when the comparison is restricted to the volunteers. In the other two groups (Medium and Low), the whole-class groups are said to have done "as well as but not better than" the skill grouped.

Year 2: Did not report the test broken into two parts. For the high-skilled students, "no impairment" is reported from whole-class grouping, but we are not told which group scored higher. For the Medium and Low groups, the whole-class groups are reported as outscoring the skill grouped, and the result is statistically significant.

Teacher-made test on the U.S. Constitution.

Year 1: Among the Low-skill students, the whole-class students scored significantly higher than the skill-grouped students. The High- and Medium-level students in whole-class groups "did as well on the Constitution" test as their peers in skill-grouped classes.

Year 2: For the High-skill students, no significant difference existed between the skill-grouped and whole-class grouped. For the Medium and Low students, the whole-class groups scored significantly and substantially higher.

For the two years, the High-skill students probably did better in skill-grouped classes while the Medium and Low students performed better in whole-class instruction. Based on this qualitative and quantitative summary, we assigned effct size –.16 for Marascuilo & McSweeney's study.

Non-cognitive findings: In the first year, whole-class students were significantly more dissatisfied with their assignments and classwork than the skill-grouped students. In the second year, the differences diminished.

* * *

8. PETERSON, R. L.

Date of publication	*Grade of students*	*Class subjects*	*Duration of experiment*
1966	7, 8	Language, History, Arithmetic	1 year

Randomization: On the basis of three standardized general aptitude tests in seventh grade, with the addition of teacher recommendations in the eighth grade, the total population was divided into three skill levels. Each skill level was divided randomly first into students to be assigned to whole-class instruction and those to skill-grouped instruction. The ultimate whole-class groups were formed by taking one-third of students from each of the three levels.

	No. of	*Skill-level sample sizes*				
Grouping	*levels*	*High*	*Med.*	*Low*	*Total*	*All grades*
Skill, 7th	3	26	24	26	76	
8th	3	27	28	26	81	157
Whole-class, 7th	3	27	24	25	76	
8th	3	28	27	29	84	160

Effect sizes* based on 8 tests in 7th grade and 9 in the 8th grade:

7th	.28	–.43	.12	
8th	.01	–.40	–.17	
Average	.14	–.42	–.02	–.10

* Positive is favorable to skill grouping, negative to whole-class instruction.

Non-cognitive findings: Among teachers, 13 of 18 preferred skill grouping. In the eighth grade among the skill grouped, the low-skill students reported a greater dislike for their section; in contrast, they had higher scores for liking school than the low-skill whole-class students.

* * *

9. VAKOS, H. N.

Date of publication	*Grade of students*	*Class subjects*	*Duration of experiment*
1969	11	American and World History	1 year

Randomization: From the eleventh-grade class of 520, a random 87 were chosen as the experimental group and 116 as the control. The same three teachers taught both sets of classes. The Iowa Test of Educational Development was used to assign three skill levels. High group was 70th percentile and higher, Medium 25th to 70th percentile, Low below 25th, using Minneapolis norms. Attrition reduced 87 to 79 and 116 to 105.

The skill-grouped class convened as whole-class instruction 60% of the time and as skill-grouped 40% of the time.

	No. of	*Skill-level sample sizes*			
Grouping	*levels*	*High*	*Med.*	*Low*	*Total*
Skill	3	25	38	16	79
Whole-class	3	43	38	24	105

*Effect sizes**	*High*	*Med.*	*Low*	*Average*
American History	–.37	–.06	–.47	–.27
World History	.57	.23	.67	.43
Average	.10	.08	.10	.08

* Positive is favorable to skill grouping, negative to whole-class instruction.
(The complete reversal of performance from one semester to the next is puzzling.)

Non-cognitive findings: Vakos did not report non-cognitive gains/differences.

* * *

10. WARDROP, J. L., ET AL.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1967	3	Math	1 semester

Randomization: Students were ranked according to the sum of twice a mathematics test score plus the score on an I.Q. test. The population was then divided into thirds to produce three levels and stratified by sex and achievement. In each group about one-third were randomly assigned to whole-class instruction.

	No. of	*Skill-level sample sizes*			
Grouping	*levels*	*High*	*Med.*	*Low*	*Total*
Skill	3	17	23	18	58
Whole-class	3				24
Effect size*		–.01	.42 (Medium and Low)		

* Positive is favorable to skill grouping, negative to whole-class instruction.

Whether because of small sample size with associated large fluctuations or possibly misidentification of groups, the low-skill group among whole-class students scored higher than the medium-skill group on both the teacher-made test and the standardized test. To guard against the consequent complications, it seems reasonable to pool the medium- and low-skill-level scores for making the comparison between skill-grouped and whole-class students for Medium and Low, and thus to report the same number for the two skill levels.

Non-cognitive findings: Not applicable.

APPENDIX 3
Two Experimental Studies Comparing Student Performance under the Joplin Plan with Performance under Whole-Class Instruction

1. MORGAN, E. F., JR., & STUCKER, G. R.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1960	5, 6	Reading	1 year

Randomization: The students were matched on two measures of reading ability, and they formed ninety matched pairs in the two grades. Those above expected grade norms are in the High group, those below in the Low group. In each pair, one student was randomly assigned to the Joplin Plan, the other to whole-class instruction. Teachers were randomly assigned to groups.

		Sample sizes				
	No. of	*5th grade*		*6th grade*		
	levels	*High*	*Low*	*High*	*Low*	
Joplin	2	27	20	27	16	
Whole-class	2	27	20	27	16	
Effect size*		.31	.38	.17	.79	Overall .41

*Positive is favorable to Joplin Plan.

Non-cognitive findings were not discussed.

* * *

2. HILLSON, M., JONES, J. C., MOORE, J. W., & VAN DEVENDER, F.; JONES, J. C., MOORE, J. W., & VAN DEVENDER, F.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1964, 1967	1, 2, 3	Reading	3 years (1960–1961, 1961–1962, 1962–1963)

Randomization: All students assigned randomly to experimental group (Joplin) and control (whole-class). In the first year, the Joplin Plan used three levels of reading skill; in the second year, six levels were used; and in the third year, nine levels.

	Sample sizes		
	Joplin	*Whole-class*	*Effect size*
At 1½ years	26	26	.41
At 3 years	27	22	.25
At 3 years Language Test of Stanford Achievement Test:			
	4.25 grade	3.98 grade	.27

The effect sizes come from averages for tests of Paragraph Meaning, Word Meaning, and Reading. The differences at the end of 1-1/2 years are statistically significant, or almost so, and those at the end of the third year are favorable to the Joplin Plan but not significant.

Non-cognitive findings: Almost all students reported, when asked, that they enjoyed reading. "Do you enjoy reading class?" "Yes." 100% of both groups. On other questions the response rates were nearly identical for the two groups.

All six teachers favored the non-graded program (Joplin).

Parents: "If I had my choice, I would favor having my child go to school in a non-graded primary organization."

	Agree	*Disagree*
Joplin parents	16	3
Control parents	3	15

This outcome is hard to interpret because "non-graded" or "graded" might not be understood in this context by a whole-class parent. Nothing is more highly "graded" than Joplin Plan students — the language may simply be confusing. What is given up under the plan is naming the grades, first grade, second grade, and so on. A student enrolled in a Joplin plan for all subjects would not have a grade but would have a level associated with each subject, such as reading level 3, arithmetic level 5, geography level 2, etc.

APPENDIX 4
Three Experimental Studies Comparing Student Performance under Within-Class Grouping with Performance under Whole-Class Instruction

1. DEWAR, J. A.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1963	6	Arithmetic	?

Randomization: Students were grouped into three subgroups. Eight teachers (and classes) were assigned at random to teach control or experimental classes of about twenty-five students each. Both types of classes were divided into three groups (High, Medium, and Low) in a similar manner. (The whole-class groups were similarly divided for purposes of analysis, not for teaching.)

	No. of	*Sample sizes*			
Grouping	*levels*	*High*	*Med.*	*Low*	*Total*
Within-class	3	28	40	30	98
Whole-class	3 (for analysis)	34	38	29	101
Effect size*		.4	.4	.6	.47

* Positive is favorable to within-class.

Non-cognitive findings: Teachers reported more and better learning by students in High-skill and Low-skill groups. They said that teaching the within-class skill groups took more organization time, but was not more difficult than whole-class instruction. Within-class grouped students said that under within-class instruction "teachers had more time to help pupils," "no need to wait for slower pupils," "full meaning," "learn more," and "it was fun."

* * *

2. SLAVIN, R. E., & KARWEIT, N. L.

Date of publication	*Class subject*	*Duration of experiment*
1985	Arithmetic	1/2 year

The within-class skill grouping was the Ability-Grouped Active Teaching (AGAT) instruction program. The MMP used as a control group in Experiment 1 was based on the Missouri Mathematics Program, whereas the additional control group in Experiment 2 was whole-class grouped with no special instruction for the teacher.

Randomization: On the basis of an initial test, each AGAT class was divided into two skill groups — 60% High and 40% Low. Teachers were to push the pace of the High group.

EXPERIMENT 1
(Grades 4, 5, and 6)

Grouping	*No. of levels*	*No. of students*	*Computation*	*Concepts and Applications*
AGAT	2	133		
MMP	1	89		
Effect size*			.74	.08

*Positive is favorable for AGAT, negative for MMP.

EXPERIMENT 2
(Grades 3, 4, and 5)

Control is an untreated control group, presumably yielding less effective teaching than MMP.

Grouping	*No. of students*	*Computation*	*Concepts and Applications*
AGAT	98		
MMP	162		
Control	106		
Effect sizes*:			
AGAT against MMP		.55	.63
AGAT against Control		.84	.73

*Positive favors AGAT

Non-cognitive findings: In neither experiment did AGAT or MMP students differ in liking math or in self concept.

* * *

3. WALLEN, N. E., & VOWLES, R. O.

Date of publication	*Grade of students*	*Class subject*	*Duration of experiment*
1960	6	Arithmetic	1 year

Two schools used a cross-over design, two teachers in each school, changing method of instruction from first semester to second.

Randomization: Students were ranked according to the arithmetic subtest of the California Achievement Tests. Scores were ranked and alternate scores were assigned to same class.

	Each semester		
	Number of students		
Grouping	*School 1*	*School 2*	
Within-class	25	31	
Whole-class	25	31	
Effect sizes*:			
non-group then grouped	–.04	.26	
grouped then non-grouped	.25	–.15	
Average effect size			.08

*Positive is favorable to within-class grouping, negative to whole-class instruction.