8. THE ROLE OF FIELD TRIALS IN EVALUATING SCHOOL PRACTICES: A RARE DESIGN

BILL NAVE, EDWARD J. MIECH, and FREDERICK MOSTELLER

"It sounds like a good idea, but does it work in practice?" Whenever educators propose reforms in schools—such as reduced class size, cooperative learning, or expanded preschool—this fundamental question about effectiveness needs answering. To find an answer, educators have turned to a repertoire of strategies, most frequently ones based on survey data. They have, however, largely neglected a powerful and persuasive research design to demonstrate program effectiveness: the randomized-controlled field trial. Widely used in other disciplines, such as medicine and public health, this design appears to be rare in U.S. evaluation research of education practices in preschool through 12th grade (pre-K-12). Because of the power of field trials to reflect results rather than intentions in evaluations of school practices and to link interventions to outcomes, this infrequent use of field trials needs to be examined.

In this article, we first define "randomized-controlled field trial" and discuss its strengths in demonstrating program effectiveness. Second, we offer several examples of randomized-controlled field trials in education in the hope that it will increase historical awareness of field trials and show how this design has contributed valuable knowledge about school practices. Third, we describe some steps that might make field trials more relevant in educational research.

DEFINING FIELD TRIALS

When we speak of field trials, we specifically refer to *randomized-controlled field trials.* Because field trials can be confused with various types of "experiments" or

D.L. Stufflebeam, G.F. Madaus and T. Kellaghan (eds.). EVALUATION MODELS.

"comparison studies," we provide a concise definition of the field trial design. In a field trial, researchers assign participants at random to control and experimental groups and then compare the results when the experimental group (or groups) receives some intervention[1] and the control group receives some other treatment. The intervention takes place in a real-world setting of practice, such as a regular classroom, and not in a more artificial setting such as a psychological laboratory. Because participants have been randomly allocated, the difference in performance between experimental and control groups can be reasonably attributed to the differential effect of the experimental treatment.

A brief example from a recent field trial in education can help illustrate its design features. In 1985, researchers in the Tennessee study on class size randomly assigned about 6,400 kindergarten students and 300 experienced teachers to one of three groups formed at each participating school: "small" classes, with 13–17 students; "regular-size" classes with 22–25 students; and regular size classes with a teacher's aide. Students then remained in their small or regular-size classes for the next four years, from kindergarten to the end of third grade.[2] Researchers compared the average reading and math performance of students in the three groups, and, based on these findings, were able to demonstrate that small class size did have a favorable effect on student achievement (Blatchford & Mortimore, 1994; Finn & Achilles, 1999, 1990; Mosteller, 1995; Mosteller, Light, & Sachs, 1996; Nye, Hedges, & Konstantopoulos, 1999; Word et al., 1990).

STRENGTHS OF FIELD TRIALS

Perhaps the greatest strength of field trials is their ability to demonstrate that a specific treatment caused certain effects. Without the random assignment of participants to experimental and control groups,[3] it can prove extremely difficult to convince others—as well as one's self—that differences in results between groups at the end of a program can be ascribed to the treatment rather than to preexisting differences in individuals in the two groups.[4]

The ability to assign effects to treatments can be especially important when dealing with small but valuable effects (or, alternatively, lack of effects or negative effects) of a program under evaluation. If a program has a huge effect on its participants, an evaluation with rigorous design may be unnecessary. For example, if people recover when given a new treatment for a disease, whereas formerly people with the disease all died in short order, then the evidence favoring the new treatment is compelling. For such large effects, the dramatic results speak for themselves and clearly seem to be the result of the intervention.

But relatively few programs produce effects large enough to meet this "slam-bang" criterion. In less dramatic circumstances, other differences between the groups might serve as rival explanations for the results of the intervention. Overwhelming effects are generally rare in social or medical interventions and are similarly rare in education interventions because factors such as family socioeconomic status and level of parental schooling have long been established as major explanatory variables for differences in student achievement.

For example, the Tennessee class size study showed an effect size of about .25 of a standard deviation for the performance of elementary students in small classes on reading and math standardized test scores, when compared with their peers in regular-size classes. This effect size translates into moving the average student, who formerly performed at the 50th percentile level, to the 60th percentile level. If the evaluators had not used a rigorous design, they would have found it difficult to state with confidence that the .25 effect size was due to differences in class size rather than to other factors such as differences between the teachers in the experimental and control groups (perhaps the group of teachers with the smaller class sizes were more talented practitioners, on average, than the teachers with the regular class sizes); differences between the students (perhaps the students in the smaller classes came from families with higher socioeconomic status and more parental education, on average, than students in the regular classes) or differences between the schools.[5]

Another strength of randomized-controlled field trials is the credibility of their findings to those both inside and outside the education community. The overall straightforwardness of field trials—the idea that several comparable groups were formed, and treatment groups received the experimental program whereas the control groups did not—can appeal to a diverse constituency, from teachers and parents to policymakers and the general public. For example, in the aftermath of the Tennessee class size study in 1989, the Tennessee state legislature allocated several million dollars to implement small K-3 classes in the 17 school districts that served communities with the lowest per-capita incomes in the state.[6]

A NOTE ON THE ETHICS OF FIELD TRIALS

Some have objected to designs that deliver new treatments to some but not all students because students assigned to the control group are denied access to the educational program under evaluation. Although we believe this is an important concern, we think some reflection on the usual state of affairs in schools places these objections in a larger context in which they lose much of their strength.

U.S. schools are generally awash in innovation: new educational ideas, programs, and reforms are constantly being implemented in schools and classrooms, often at the same time (Cuban, 1990; Elmore, 1996). Advocates of these innovations bring intelligence and good intentions to this task of improving schools and usually have a theory about how a particular innovation, once implemented, will benefit students and educators. These innovations, however, frequently play out differently in practice than originally predicted.[7] Some students may benefit from the effects of the innovations, while others may not. Further, because the innovations usually occur without systematic evaluation to gauge their relative effectiveness, policymakers have no sound basis for deciding whether to expand, modify, or scrap the new programs.

With field trials, by contrast, researchers can evaluate the effects of education programs and provide compelling evidence either to support the broad-scale implementation of innovations that prove successful, or to avoid false steps and wasted

resources that would result from implementing ineffective innovations. In this way, rigorous designs convey serious respect for the teaching and learning process and for the risks students run when participating in new programs, just as Tennessee did.[8]

Furthermore, policymakers can use research findings from field trials to inform large-scale school improvement efforts elsewhere. At least 30 states have initiated class size reduction measures since the Tennessee study, and California alone has invested approximately $3 billion in reducing class sizes in the early grades (Finn & Achilles, 1999, p. 104). Overseas, the Republic of Ireland has implemented its own class size reduction initiative in some of its more economically-depressed regions (Kellaghan, Weir, Ó hUallacháin, & Morgan, 1995).

NOTABLE FIELD TRIALS OF PRE-K-12 EDUCATION

Readers might wish to ask themselves this question: Off the top of your head, if you were asked to name some well-known field trials in U.S. education, what comes to mind? At first, we had a difficult time answering the question ourselves. In subsequent review of the literature, we identified seven such studies that we felt might interest the larger education community, and we describe them in this section. We make no claim for the defmitiveness or comprehensiveness of this list, but rather offer it as an effort to present a useful, thought-provoking collection of randonmized-controlled field trials in education.

We present the field trials in an order roughly corresponding to the strength and direction of the studies' findings (from strong positive effects to zero effects to results still pending). For each we describe the study, summarize its findings, provide a measure of its influence,[9] note its possible policy implications, and comment on policy decisions it may have influenced.

Tennessee Class Size Study (1985–1989)

The randomized-controlled field trial known as Project STAR (Student-Teacher Achievement Ratio), conducted in Tennessee in the mid-1980s, is probably the largest, most important field trial in public schools ever funded by a state. Project STAR studied the effects of small class size on student achievement in kindergarten and grades one, two, and three, and involved about 80 public elementary schools throughout Tennessee.

The origins of Project STAR date to the early 1980s when Lamar Alexander, then governor of Tennessee, sought to improve public schools in his state. A modest study in neighboring Indiana, Project PrimeTime, suggested that smaller class sizes in kindergarten through third grade enhanced student achievement in school. This two-year study was interrupted after two semesters because Indiana was so impressed by the gains in student performance that they decided to implement smaller classes statewide immediately. Prior to committing large sums of money for the purpose of reducing class sizes, Governor Alexander and the Tennessee Legislature agreed to fund a four-year $12-million randomized-controlled field trial to determine the effects of reduced class size and of teacher's aides on student achievement in the

lower grades. The Tennessee legislature required that the study include students from inner-city, suburban, urban, and rural areas; schools across the state were invited to participate in the study (Mosteller, 1995; Word et al., 1990). The introductory sections of this chapter described Project STAR's sample size, its class size intervention, and the study's findings.

A follow-up study to Project STAR, begun in 1989, asked if the differential achievement effects of small classes continued after all students participating in the field trial moved to regular-size classes in the fourth grade. This observational study, known as the "Lasting Benefits Study," found that improved student achievement continued through at least the eighth grade for students who were in the small classes for kindergarten and grades one, two, and three during Project STAR (Achilles et al., 1993; Nye, Hedges, & Konstantopoulos, 1999).

Based on the results of Project STAR, the Tennessee legislature voted in 1989 to allocate millions of dollars to institute small class sizes in kindergarten and grades one, two, and three in the 17 school districts in Tennessee with the lowest per-capita incomes in the state. This initiative, known as "Project Challenge," has also yielded impressive results in follow-up observational studies: the average end-of-year rank of second-grade reading scores for students in the 17 districts rose from 99th in 1990 (out of a total of 138 school districts in Tennessee) to 78th in 1993; the average end-of-year rank of math scores for the same group of students during the same time period rose from 85th rank to 56th (Achilles, Nye & Zaharias, 1995).

A search in the Social Science Citation Index (SSCI) for citations of the primary articles detailing the Tennessee study yielded a total of 80 citations. This relatively modest number of citations in the scholarly literature suggests that the results of the study may not yet be widely known in the educational research community. This state of affairs is apparently changing, however. In a recent special issue on class size findings in *Educational Evaluation and Policy Analysis,* a quarterly journal published by the American Educational Research Association, half of the featured articles were about the Tennessee study (Finn & Achilles, 1999; Hanushek, 1999; Nye, Hedges, & Konstantopoulos, 1999; Ritter & Boruch, 1999). In the introduction to the issue, David Grissmer of the RAND Corporation noted the growing influence of the Tennessee study on the policymaking community:

> . . . the Tennessee experiment has had significant influence among policymakers. . . . Although the Tennessee results were known as early as 1990, they did not receive much attention from the research or national policymaking community until years later. Initially, the results seemed to be treated as simply one more set of findings—among scores of studies done on class size. . . . But the results of the Tennessee study are increasingly being interpreted by many as "definitive" evidence that supplants the scores of studies using non-experimental data. (Grissmer, 1999, p. 93)

The High/Scope Perry Preschool Study (1962–1965)

The High/Scope Perry Preschool Project was a randomized-controlled field trial begun in the 1960s. It investigated the short- and long-term effects of an intensive,

high-quality preschool program for children from economically disadvantaged backgrounds (Barnett, 1985, 1993, 1995, 1996, 1998; Schweinhart et al., 1993; Schweinhart & Weikart, 1997; Zigler & Styfco, 1994).

Several hypotheses served as the foundation for the study: that a good preschool program could help young children who were at high risk of school failure to develop the cognitive skills needed to succeed in school and thus graduate from high school; that preparation for school could be linked to success in school; "that good preschool programs can help children in poverty make a better start in their transition from home to community and thereby set more of them on paths to becoming economically self-sufficient, socially responsible adults"; and that success in school could be linked to success in the "real world" of jobs, families, and community (Schweinhart et al., 1993, pp. 3–7). Another goal for the High/Scope Perry Preschool Project—"too bold at the time to be framed as a hypothesis"—was that participants would ultimately be less likely to be involved in the criminal justice system because they were more successful in school (Schweinhart et al., 1993, p. 7).

The sample size for the study was modest. From 1962 through 1965, 123 African-American children in Ypsilanti, Michigan participated in five waves, with an average of approximately 25 children per wave.[10] These three-year-olds (except for "Wave Zero," which involved four-year-olds) were identified as living in poverty and assessed to be at high risk of school failure. Children were randomly assigned to a treatment group, which received the Perry Preschool program for two years (except for the four-year-olds in "Wave Zero," which received only one year of preschool) and a control group, which did not receive any preschool program.

The Perry Preschool program consisted of a daily 2 1/2 hour classroom session for children on weekday mornings, and a weekly 1 1/2 hour home visit to each mother and child on weekday afternoons. The curriculum heavily emphasized active learning, in which "children plan, or express their intentions; carry out, or generate, their play experiences; and reflect on their accomplishments" (Schweinhart et al., 1993, p. 227).

A striking and important feature of the Perry Preschool field trial has been its 30-year longitudinal reach with little attrition from the original groups of participants. Researchers collected data on the 123 individuals in the treatment and control groups annually from ages 3 through 11, then at ages 14–15,19, and 27, and reported the results of their data analyses after each of these phases (Barnett, 1993, 1996; Schweinhart et al., 1993; Schweinhart & Weikart, 1997). A number of assessment instruments and data-gathering techniques were used at various times throughout the study, including interviews; cognitive, performance, and behavior instruments; and analyses of public and private records from sources such as schools, police departments, courts, and social services."

A wide variety of long-term benefits were associated with participation in the Perry Preschool program. In educational benefits, students in the preschool group had significantly higher average IQ scores than students in the control group from

the end of the first year of the preschool to the end of the first grade, significantly higher school achievement at age 14, and significantly higher general literacy scores at age 19. They had a significantly higher level of schooling completed, with an average of 71 percent completing 12th grade or higher, as compared to 54 percent of the students in the control group. In addition, students in the preschool group spent significantly fewer years in special education in programs for "educable mental impairment" during their school careers, with 15 percent of the preschool group and 34 percent of the control group spending a year or more in one of these programs (Barnett, 1993; Schweinhart et al., 1993; Schweinhart &Weikart, 1997).

The study also showed lasting economic and social benefits from participation. In the 1990s, adults who had attended Perry Preschool in the 1960s had significantly higher monthly earnings at age 27 than students in the control group, with 29 percent of the former vs. 7 percent of the latter earning $2,000 or more per month (Schweinhart, 1993; Schweinhart & Weikart, 1997).

This economic self-sufficiency also translated into a significantly lower percentage of adults in the preschool group in receipt of social services at some time over the previous decade (59 percent) as compared with adults in the control group (80 percent). Adults in the preschool group also had significantly fewer arrests by age 27, with 7 percent of the program group and 35 percent of the control group having five or more arrests (Barnett, 1993; Schweinhart et al., 1993; Schweinhart & Weikart, 1997).

Based on their overall findings, the researchers had ample evidence to support the basic hypotheses they formulated at the beginning of the study in the areas of educational performance, delinquency and crime, economic status, family formation, childrearing, and health. (Schweinhart et al., 1993, p. xviii; see also Barnett, 1993, 1995, 1996, 1998; Schweinhart & Weikart, 1997). The results of the Perry Preschool randomized-controlled field trial have supported policymakers in their decisions to fund preschool programs for disadvantaged children in the United States. The eight primary High/Scope publications reporting on the results of the Perry Preschool Experiment since 1967 have over 500 citations in the SSCI, making it one of the best-known randomized-controlled field trials in education.

Pygmalion in the Classroom (1964–1966)

The field trial that came to be known as *Pygmalion in the Classroom* examined teacher expectancy effects in an elementary school identified as "Oak Park School" (Rosenthal & Jacobson, 1968). The study involved over 500 students enrolled in kindergarten through fifth grade,[12] and was "designed specifically to test the proposition that within a given classroom those children from whom the teacher expected greater intellectual growth would show such greater growth" (Rosenthal & Jacobson, 1968, p. 61). In other words, the *Pygmalion* study investigated whether teachers' perceptions of student ability could actually lead to changes in a child's cognitive performance. The field trial also examined teacher expectancy effects by grade level, track level, gender, and minority group status.

In the spring of 1964, all students in grades kindergarten through 5 in Oak Park School were given the "Harvard Test of Inflected Acquisition." Teachers were led to believe that the test was in its final stage of validity testing and that it was designed to predict academic "spurting" or "blooming." In reality, the test was Flanagan's 1960 Tests of General Ability (TOGA), a relatively nonverbal test of intelligence available in both Spanish and English (Buros, 1953). The test was chosen for a variety of reasons: it was unlikely that any teachers at "Oak Park School" had seen it; Oak Park School had many bilingual students with poor English skills, and the test did not rely heavily upon school-acquired skills such as reading, writing, and arithmetic; and it was group-administered.

The following school year, Oak Park School teachers were given a list of the students in their class who were likely to bloom academically because these students supposedly had scored among the top 20 percent on the TOGA. In reality, however, the researchers selected these "late bloomers" using a table of random numbers. These students were distributed among 18 of the school's teachers, one in each track (high, middle, low) for each of the grades 1–6. Lists varied from 1 to 9 students in each class, and varied from 33 to 66 percent female (on lists of more than one student).

To measure expectancy effects, posttests were given one semester, one year, and two years after the initial administration of the TOGA. The first two posttests were administered by the teachers, who had been given reason to expect late blooming on the part of some of their students and who had also been told that these additional tests were part of a further attempt by the researchers to predict late-blooming students.

Two different scorers independently scored the TOGA; neither scorer knew whether children were in the treatment group or in the control group. Statistically significant gains in IQ points were found for the treatment group students in grades 1 and 2, and for girls in the middle track.[13] No main effects were found to be associated with any of the three academic tracks at the school.

Pygmalion in the Classroom has proven to be one of the more controversial randomized-controlled field trials in educational research. Several scholarly studies, such as *Pygmalion Reconsidered* by Janet Elashoff and Richard Snow (1971), have critiqued the design and implementation of the study in considerable detail. The controversy and discussion of the merits of the original study, and the many that have followed it, have continued over the last 35 years (see, for example, Rosenthal, 1994 and Spitz, 1999).

Nevertheless, *Pygmalion in the Classroom* has also proven to be one of the more influential field trials in educational research. The original study has been cited more than 1,400 times since its publication in 1968 making it perhaps the most widely-known field trial in U.S. education. In addition, over 400 studies have been carried out to test or extend the findings of expectancy effects.[14] Seven meta-analyses carried out by Robert Rosenthal and others between 1968 and 1990 consistently find that 35 to 40 percent of these studies result in statistically significant effects.

We return to the subject of teacher expectancy later in this chapter when we discuss the study of the effects of standardized testing, in which the researchers performed an extensive analysis of expectancy effects as part of their national randomized-controlled field trial in Ireland. In terms of policy influence in the United States, the importance of "teacher expectation" has become a truism in the 1990s. A widely-held belief is that good teachers have "high expectations" for their students.

The Carolina Abecedarian Study (1972–1985)

Noting that ". . . there are actually few scientifically rigorous studies of the efficacy of early educational programs, with subjects randomly assigned to treatment and control conditions, and periodic long-term assessment of the outcomes" (Campbell & Ramey, 1995, p. 744), the Abecedarian researchers set out in 1972 to determine, among other things, whether an educational intervention to improve education outcomes for children born into poverty was more effective at preschool or during early elementary school.

The researchers selected a set of healthy infants (N = 109) born to poor families living in a small town in the southern U.S. Half the infants (N = 55) were randomly assigned to a specially designed five-year preschool program that extended from the first year of life until the time to enter public kindergarten, whereas the other half (N = 54) were randomly assigned to a control group. At the end of five years, the preschool group and the no-preschool control group were randomly split again. Half of the preschool group, as well as half of the no-preschool group, were randomly assigned to a three-year school-based intervention covering grades K, 1, and 2; the other half of each group received no school-age intervention. Thus, there were a total of four groups in the study: one with eight years of intervention (5-year preschool plus 3-year K-2 school-based intervention, N = 25), one with five years (preschool only, N = 22), one with three years (K-2 only, N = 24), and one with no educational intervention over the eight-year period (N = 22).[15]

Four cohorts, with an average of 28 infants per cohort, entered the study between 1972 and 1977. The researchers assigned each infant a risk score based on a 13-factor risk index, matched them on the basis of the score, then randomly assigned one of each pair to the experimental preschool group and the other to the preschool control group. Upon entry to kindergarten, children within each group were matched on the basis of their 48-month IQ score, then randomly assigned to the school-age intervention or the school-age control group.

The preschool was a full-day, year-round program with a caregiver-to-infant ratio of 1:3. The program's custom-designed curriculum addressed four major domains: cognitive and fine motor development, social and self-help skills, language, and gross motor skills. As children moved into the toddler and preschool stages, the caregiver-to-child ratios increased gradually to 1:6. The preschool included centers for art, housekeeping, blocks, language, literacy, and fine motor manipulatives. The language program was integrated throughout the day's activities, emphasizing pragmatic interactive features of adult-child language.[16]

The school-age intervention focused on increasing parent involvement to enhance their children's academic development. Each family in the experimental group was assigned a home/school resource teacher (HST) for the first three years their child attended public school. The HST served as a liaison, working with both parents and teachers, providing families with learning activities designed specifically for each child to support his/her work on the reading and math being taught at school.[17] Parents were encouraged to do these learning activities with their children for at least 15 minutes a day. The HST also functioned in some ways like a social worker, referring families to various agencies for services as needed.

The study found that children in the preschool treatment group fared better in several ways than students who had been in the preschool control group. The average advantage in IQ for the preschool treatment group was 8.8 points (16.4 points at age 36 months, 4.5 points at age 8, 4.6 points at age 15). Students in the preschool treatment group scored significantly higher at age 15 in reading and math than students in the preschool control group. Finally, "Through 10 years in school, children who had the Abecedarian preschool treatment made better school progress, in terms of fewer retentions in grade and fewer assignments to special education programs, than those in the preschool control groups" (Campbell & Ramey, 1995, p. 761).

The school-age portion of the treatment produced no significant effects by itself, leading the investigators to conclude that

> the value of providing only a supplemental program in the primary grades of public school appears doubtful, being, by itself, not associated with greatly enhanced academic outcomes. Even though it is easier to provide supplemental services for children once they are in school, those who plan interventions for poor children should be aware that elementary school programs may have less impact on the children's academic performance than would programs begun earlier in the life span. (Campbell & Ramey, 1995, p. 769)

The Abecedarian researchers continue to collect data as the study participants reach the age of 21, and plan to evaluate outcomes "across the full developmental span from infancy to young adulthood" (p. 769).

Taken together with earlier published reports of the Abecedarian study (Barnett, 1995; Campbell & Ramey, 1994; Ramey & Campbell, 1984, 1991; Ramey & Smith, 1977), the SSCI lists 125 citations. Although this study appears to be less well-known than the Perry study, we believe it is noteworthy because of its longitudinal follow-up (like the Perry study) and because of its attempt to discover the relative efficacy of preschool versus in-school educational interventions for disadvantaged children.

Harvard Project Physics (1967–1968)

Harvard Project Physics (HPP) was a national curriculum development effort designed to reverse the precipitous decline in the percentage of high school students enrolling in physics courses by making the course more engaging to students, especially to those not planning careers in math, science, or technology (Bottoms, 1977). The project was co-sponsored by the Carnegie Corporation, the National

Science Foundation, the Sloan Foundation, Harvard University, and the U.S. Office of Education.

Beginning in 1967, researchers conducted a year-long randomized-controlled field trial to evaluate the results of this curriculum development. Investigators randomly selected a pool of high school physics teachers from a list of most U.S. physics teachers. Teachers from this pool were invited to participate in the study, and 53 were able to do so. These were randomly assigned to the experimental group (N = 34), which received a six-week summer course on how to teach the HPP curriculum, or the control group (N = 19), which attended a two-day session at Harvard hosted by university physicists, who asked control group teachers to teach their regular physics courses and also emphasized to them the importance of their participating in the experiment.[18]

The achievement and attitudes of students in the physics classes of these two groups of teachers were then compared at the end of the academic year. Students in the HPP classrooms reported much greater satisfaction and interest in physics than their counterparts in the control group. No significant differences were found between the students in the HPP sections and the students in the control group in terms of achievement in physics (Welch & Walberg, 1972). It should be noted, however, that the Project's primary stated goal was to increase student enrollment in physics courses, not to increase physics achievement, and by this criterion, evaluators considered the Project a success.

We think the Harvard Project Physics study is notable because it represents an early example of a curriculum project evaluation designed as a national randomized-controlled field trial. Although we found many articles discussing the project itself, the brief article reporting the HPP field trial evaluation has been cited only 21 times in the SSCI. Lee Cronbach (1982) examined the HPP study in some detail as one of three examples used to illustrate field trials as evaluation tools. Using HPP's unpublished final evaluation report, Cronbach analyzed in detail the strengths and weaknesses of the study's design and analysis of results.

The Career Academies Study (1992–2003)

The Manpower Demonstration Research Corporation (MDRC) began this field trial in 1992 to examine the effects of "career academies" on high school students' academic achievement, progress towards graduation, and preparation for postsecondary education and employment. Career academies are specialized public high schools that combine academic and vocational instruction and provide work-based internships as a way to prepare students for college, employment, or both. Each of the nine[19] career academies participating in the MDRC study has a career theme, such as aerospace technology, business, electronics, health, or public service. The nine career academies are scattered throughout the country[20] (Kemple & Rock, 1996; Kemple & Snipes, 2000).

Over a three-year period, beginning with the 1992–93 school year, about 1,700 eighth- and ninth-grade students participated in a lottery in which they were randomly assigned to a "program group" or a "control group." Students in the program

group (N = 959) enrolled in a career academy, and students in the control group (N = 805) enrolled in their traditional high school (or participated in another option offered by the school district). Data used in the study included school records on attendance, achievement, course-taking patterns, and progress through high school (Kemple & Rock, 1996; Kemple & Snipes, 2000).

In 2000, MDRC released its first report on the study that included an analysis of student performance data (Kemple & Snipes, 2000). The report, assessing the progress of the student cohort from 8th and 9th grade through the end of 12th grade, found that "high-risk" students in the Career Academies had substantially reduced dropout rates along with improved attendance, increased academic course-taking, and increased likelihood of earning enough credits to graduate on time when compared with their high-risk counterparts in the control group.[21] "Low-risk" students in Career Academies had an increased likelihood of graduating on time compared to the corresponding subgroup in the control group. On average, all students in the Career Academies received more interpersonal support at school and participated more in career awareness and work-based learning activities than students in the control group (Kemple & Snipes, 2000).

However, of the 490 students (out of the study total of 1,764, or 28 percent) who completed standardized math and reading tests[22] at the end of their 12th grade, no significant differences were found in math or reading performance between the students in the Career Academies and their counterparts in the control group. Furthermore, when all students in the study were averaged together, the Career Academies showed only small reductions in dropout rates and slight increases in other measures of school engagement (Kemple & Snipes, 2000).

The Career Academies field trial will continue through 2003 to follow students through postsecondary education and employment to evaluate the impact of career academies on future educational and economic prospects.

The Career Academies study is a notable example of a field trial in educational research for several reasons. First, the scope of the study—with 9 sites, about 1,800 students, and a time span of 10 years—is substantial. Second, the experience of the Manpower Demonstration Research Corporation in carrying out the study may be of considerable value to the education research community. MDRC, based in New York City, has a long-established reputation for designing rigorous investigations to study the economic impact of programs intended to benefit disadvantaged populations, including youth from low-income families. The Career Academies study is MDRC's first major education evaluation in its 25-year history (Kemple & Rock, 1996; Kemple & Snipes, 2000). Third, the Career Academies study is funded by a consortium of seventeen private foundations in addition to the U.S. Department of Education and the U.S. Department of Labor, a private/public funding strategy similar to that which supported the Harvard Project Physics evaluation. Fifth, the Career Academies study continues to look to the future. As a randomized-controlled field trial, it provides an opportunity for the education community to look forward with anticipation to the results of a study that bears directly on an important issue for pre-K-12 practice: the subject of school-to-work transition.

Table 1. Description of control and treatment groups in the standardized test study in Ireland.

First control group	Students not given any standardized tests
Second control group	Students in grades 2–6 were given standardized tests of ability and achievement, but no feedback on test performance was provided
Treatment groups	Students in grades 2–6 took standardized tests and teachers received standardized scores and percentile ranks of all students; in some treatment groups, teachers also received additional diagnostic information on individual students

The Effects of Standardized Testing (1973–1977)

This study, conducted over a four-year period in the mid-1970s, examined the effects of standardized testing and of the use of test data on school organization, teacher attitudes and practices, and student attitudes and achievement. The study took place in elementary schools in the Republic of Ireland, which did not have a prior tradition of using standardized tests.

The researchers stratified the approximately 3,400 elementary schools in the country by pupil composition (all male, all female, or mixed) and location (city, town, or rural); schools were randomly selected within each stratum. For each school selected, four additional schools matched by size (number of teachers) and administration (lay or religious) were randomly selected. Each school within this matched set of five schools was randomly assigned to one of several treatment or control groups (see Table 1). The final sample of 35 sets of 5 matched schools yielded a total sample of 175 schools.

In the treatment groups, there was considerable planned between-group variation in terms of which students took the standardized tests, whether the students took norm-referenced or criterion-referenced tests, and what types of information was given to teachers (Kellaghan, Madaus, & Airasian, 1982).

The study found that standardized testing had little impact on schools. Admission practices and report cards were unchanged, communication practices remained the same, grouping decisions were largely unaffected, and decisions about students with learning difficulties were not altered. The researchers concluded that their findings "provide no evidence to support the position that standardized testing, when based in classrooms under the control of teachers, differs in kind in its effects from any other evaluative procedure available to the teacher" (Kellaghan, Madaus & Airasian, 1982, p. 261).

The researchers were able to use the data generated in this large national study to examine the role of teacher expectancy effects on student achievement, the topic investigated in *Pygmalion in the Classroom.* Rather than changing teacher expectations by identifying so-called late bloomers as was done in the *Pygmalion* study, the study looked for evidence of expectancy effects in natural classroom settings by analyzing changes over time in the relationships between student test scores and teacher expectations or teacher perceptions of students' abilities and achievement potential.

It was argued that if teacher expectancy effects occurred, these effects should be evident both in the control groups of teachers who received no test information about their students and in the treatment groups of teachers who did receive that test information. They found that

> when test information was made available to teachers, their subsequent ratings of the pupils' intelligence and scholastic achievement tended to move into line with that information. . . . If, on the other hand, test information was not available to the teachers, pupils' subsequent test performance tended to move into line with initial teacher perceptions of their intelligence and achievement, in comparison with the group that received test information. . . . Thus, an expectancy process seems to have been operating in classrooms, regardless of whether or not standardized norm-reference test information was provided to teachers. (Kellaghan, Madaus, & Airasian, 1982, p. 199)

This randomized-controlled field trial is notable for its scope. Its random sample of elementary schools is meant to generalize to an entire country, and it tackles one of the largest issues in educational evaluation: the effects of standardized testing. We note that arguably the most ambitious field trial discussed in this chapter is among the least well-known. The study, carried out in the 1970s, has been cited fewer than 25 times in the SSCI, suggesting that it remains largely overlooked by educational researchers.

CONCLUSION

Each school year many new programs and innovations are introduced into U.S. classrooms, affecting the lives of millions of students, teachers, parents, and administrators. Policymakers and the general public need good evaluations of these programs in practice to make informed decisions about the deployment of school resources to benefit children. Field trials afford special advantages in establishing the benefits or shortcomings of educational interventions.

We think that researchers could conduct field trials in education more often if three factors could be aligned: resources, expertise, and leadership. Because constitutional authority for public education in the U.S. is vested in the states, a large portion of state budgets flow to public education. State-level resources made Project STAR possible in Tennessee, and many opportunities exist for individual states and groups of states to use their organizational and fiscal resources to launch field trials. Likewise, consortia of cities, foundations, or universities might find it practical and economical to study classroom practices with field trials. The federal government might also contribute resources to such trials; the Career Academy study, for example, was funded by a group of 17 private foundations in addition to the U.S. Department of Education and the U.S. Department of Labor.

In terms of expertise, we believe that ample capacity exists. In our estimation, there are at least a dozen organizations and centers around the U.S. that have the technical knowledge and experience to assist in the design and execution of a field trial in education. It takes leadership, however, to couple organizational resources

with expertise. This leadership on behalf of field trials could come from many different places, from elected officials to public administrators to concerned citizens. In Tennessee, for example, a key actor behind Project STAR was educator Helen Bain. Bain not only carried out a pilot experiment in Tennessee on class size before Project STAR, but also visited and discussed this proposal with every Tennessee state legislator and gained approval from the Tennessee Education Association, the state teachers' union. Bain's leadership was crucial (Ritter & Boruch, 1999, pp. 117, 120-121).

Field trials appear to be tools that are rarely used in the set of evaluation strategies for education. We hope that this chapter will raise awareness of their value, and that members of the education community and general public will consider using this design as part of a research strategy to identify effective educational practices. If leadership can bring together resources and expertise to rework the role of field trials in education, trials could help improve student learning by focusing on results and revealing progress on the ground.

NOTES

The preparation of this paper 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.

We wish to thank Charles Abelmann, John D. Emerson, Howard Hiatt, Nathan Keyfitz, Richard Light, George Madaus, Lincoln Moses, Marjorie Olson, Igor Perisic, Jason Sachs, John Tyler, and Cleo Youtz for their thoughtful and helpful comments on earlier drafts of the article. We thank Julius Richmond for historical background on the Perry Preschool Study and the early years of the federal Head Start program and Thomas Kellaghan for information on the class size initiative in Ireland.

1. In education, this experimental treatment is typically a modification of an existing program or a completely new program intended to improve the outcome obtained under usual conditions.

2. Some changes to the original design were made during the four-year course of Project STAR. For a fuller discussion of these modifications, see Mosteller (1995).

3. One way to form the groups of children is to assign them randomly as individuals into two or more different groups. Then the unit of analysis is the child. Sometimes, however, this is not convenient or feasible. Another approach might deal with classrooms or even schools. A collection of classrooms or of schools might be assigned randomly to one or another treatment. In this case, the unit of analysis would be the classroom or the school. For a more detailed look at randomization and the design and use of field trials in evaluation of educational and other social programs, see Boruch (1997), Cook and Campbell (1979), and Cronbach (1982).

4. Not only is it valuable to detect the actual benefits of a particular treatment, but it is also worth knowing that a treatment yields little or no benefit. Otherwise the treatment in question might be continued, giving the mistaken impression that a problem has been solved when it has not. Continuing treatments with little or no benefit can be costly in other ways. For example, after careful appraisal of research results, the medical community no longer considers radical mastectomy the treatment of first choice for breast cancer.

5. Another example of a field trial detecting a small but valuable effect is the 1954 Salk study, a landmark randomized-controlled field trial in medicine that tested the effectiveness of a new vaccine for polio with about 750,000 children in the first through third grades. The Salk study showed a drop in the incidence of cases of polio from 57 per hundred thousand (0.057%) of the non-vaccinated control group to 16 per hundred thousand (0.016%) of the vaccinated group An effect of this size, though tiny (less than 1/l0th of 1%), benefited thousands of children by verifying the efficacy of the vaccine (Meier, 1972).

6. Researchers can also use field trials to investigate specific between-group differences. In one field trial on the effects of tracking, for example, a researcher looked at class participation in question-and-answer sessions in non-tracked and tracked classrooms. In non-tracked classrooms, the researcher found that the more skilled students dominated the time, whereas in tracked classes the less skilled students were able to participate equally (Drews, 1963). Replication of this study would be valuable.

7. The field of education does not seem to have an analysis of innovations that succeed versus those that fail. Surgery provides an illustration of such an analysis: in 13 innovations in surgery intended to improve the patients' outcomes to their primary disease, 6 showed improvement over standard treatment and 7 did not. In 24 innovations intended to improve the patients' recovery from the surgery, 15 showed improvement over standard therapy, 8 showed worse performance, and 1 was a tie. In each instance, as in education, the innovator was confident that the innovation would be a success (Bunker, Barnes, & Mosteller, 1977, pp. 132–133).

8. For a detailed discussion of the ethics of conducting field trials, see Boruch (1997).

9. We use the number of citations in the Social Science Citation Index (SSCI) as of February 2000 as a proxy for the study's influence, reasoning that many citations in the scholarly literature suggest a broader influence than fewer citations might.

10. Children entered the study annually in five waves. In 1962, the first year of the study, there was a "Wave Zero" and a "Wave 1."Wave Zero involved only four-year-olds, where children in the treatment group received one year of preschool. In Wave 1, a group of three-year-olds randomly assigned to the treatment group received two years of preschool. The process for Wave 1 was repeated for Wave 2 in 1963, Wave 3 in 1964, and finally with Wave 4 in 1965 (Barnett, 1985).

11. These included initial parent interview, interviews with youths and parents at age 15, interview at age 19, case-study interview at age 21, interview at age 27, the Stanford Binet Intelligence Scale, the Adapted Leiter International Performance Scale, the Illinois Test of Psycholinguistic Abilities, the Peabody Picture Vocabulary test, the Wechsler Intelligence Scale for Children, California Achievement Tests, the Adult APL Survey, the Pupil Behavior Inventory, and the Ypsilanti Rating Scale.

12. Over 500 students initially took the IQ test that was the foundation for the study. Fewer than 400 took the first year's retests, and fewer than 300 took the two-year follow-up test. Reasons for the declining numbers were students moving away, illness at time of testing, and the sixth graders (first retest) and fifth graders (second year retest) having moved-into the junior high school.

13. These findings should be interpreted in light of the issue of multiple comparisons. When many comparisons are made, some observed differences will stand out as a result of chance fluctuations. For example, imagine that 20 independent comparisons are made, and that the 5% level is used as a criterion for considering a difference as "significant." In this case, on the average, one of the twenty comparisons will stand out by chance alone. Since grade level, track, gender, and minority status are specified in the *Pygmalion* study, it is likely that several comparisons were made; hence, due to the multiple comparison problem, the findings may not be as "significant" as the 5% level being used suggests.

14. Expectancy effect experiments have been carried out in studies of physical fitness, psychotherapy, nursing homes, the workplace, ordinary social situations, courtrooms, psychosocial judgments, inkblot tests, and reaction time, among others (Rosenthal, 1994).

15. At the beginning of the experiment, 122 families were considered eligible. Attrition for various reasons yielded a base sample of 111, with 93 finally fully eligible to be placed into one of the four cells of the experiment at the time of analysis after the completion of the school-age intervention. The researchers note that the subjects "lost to attrition do not differ from the others on any entry-level demographic characteristics" (Campbell & Ramey, 1995, p. 749).

16. Assessments used during the preschool portion on the study included the Bayley Scales of Infant Development for the infants, and the Stanford-Bmet Intelligence Scale, Form LM (at ages 24, 36, and 48 months) and the McCarthy Scales of Children's Abilities (at ages 42 and 54 months) for the preschoolers.

17. Assessments used during the school-age portion of the study included the Wechsler Preschool and Primary Scale of Intelligence (at age 5), the Wechsler Intelligence Scale for Children-Revised (at age 6.5 and again at end of treatment), the Peabody Individual Achievement Test (fall and spring of first

two years of school), the Woodcock-Johnson Psycho-Educational Battery, Part 2: Tests of Academic Achievement (fall and spring third year of school), and the Classroom Behavior Inventory (each of first three years of school).

18. The researchers brought the control group teachers to campus for the two-day meeting in an effort to avoid the so-called Hawthorne Effect that might result if only the treatment group teachers received the special attention of time on a university campus. The researchers didn't comment in the cited report, however, on the potential differential impact of six weeks of classes for the experimental group vs. the two day visit for the control group, quite apart from the impact of the curriculum itself.

19. One of the original ten career academies disbanded after two years.

20. Four of the career academies are in California, two are in Florida, and one each is located in Maryland, Pennsylvania, Texas, and Washington, DC. Each career academy in the MDRC study is a "school-within-a-school," meaning that the specialized school is physically housed in a traditional high school building, though the program is separate from the rest of the high school. The career academies are relatively small, and generally have 30 to 60 students per grade in grades 9–12 or 10–12.

21. There were 474 "high-risk" students in the study out of a total of 1,764 participants (27%). Students were identified as "high-risk" based on baseline risk characteristics including low attendance rates, low number of credits earned by 9th grade, low grade-point averages, age at grade 9, number of schools attended since 1st grade, and having a sibling who dropped out of school.

22. The test consisted of the math and reading sections of the National Educational Longitudinal Survey of 1988 (NELS: 88) Follow-Up Study.