RANDOMIZE GROUPS, NOT INDIVIDUALS

A Strategy for Improving Early Childhood Programs

ROBERT G. ST.PIERRE
STP Associates, Breckenridge, Colorado
PETER H. ROSSI
University of Massachusetts, Amherst

Until the past few years, our nation's approach to designing federal programs for preschool-age children lacked coherence and paid little attention to what had worked (and not worked) in the past. In this article, the authors propose that credible information useful for designing effective programs will require the ongoing, systematic development and evaluation of alternative approaches for the improvement of large-scale early childhood programs. The research should place greater reliance on experiments in which existing groups of individuals, such as intact classes or preschool agencies, are randomly assigned to implement competing early education programs or program components. Randomizing groups, rather than individual children, changes the research question from "What works?" to "What works better?" yielding more useful information than is currently available about which preschool approaches ought to be strongly embedded in our nation's social policy.

Keywords: *child care quality; group randomization; study evaluation*

The central message of this article is a proposal to continue the restructuring of the nation's approach to conducting federal research on preschool education and child care programs. The proposed strategy is more likely than past approaches to lead to improvements in those programs.

Preschool education receives major attention in this article, both out of a desire to simplify the presentation and as a result of the blurred boundaries between child care and preschool education. Child care is a necessity for

---

AUTHORS' NOTE: *This article has benefited from comments by Douglas Besharov, Rebecca Maynard, Howard Rolston, Robert Boruch, and several others. Whatever errors remain are entirely the work of the authors. Preparation of the article was supported, in part, by funding from Abt Associates, Inc.*

EVALUATION REVIEW, Vol. 30 No. 5, October 2006 656-685
DOI: 10.1177/0193841X06291533

many households with preschool children. Any program that cares for a child for a significant number of hours on a day-to-day basis is providing a child care service, even if the program is labeled as educational; likewise, child care programs, especially those that are center-based, often provide educational experiences. Moreover, center-based child care and preschool education are structurally similar in many ways. Accordingly, the research strategy proposed in this article for federal preschool programs is applicable, with minor modifications, to federal child care programs.

This article starts by identifying some of the kinds of knowledge that would be obtained from the recommended approach. Then it reviews the origins of public policy concern for preschool programs and presents a critique of research over the past four decades. It next considers how reliance on randomizing individuals needs to be augmented or even replaced by an emphasis on randomization of groups. Finally, it presents specific plans for a more productive federal research strategy.

WHAT MIGHT BE LEARNED BY RANDOMIZING GROUPS?

The proposed approach calls for research that focuses on the question "What works better?" instead of the more traditional "Does it work?" Such a focus necessarily leads to consideration of experiments that rely on random assignment of groups rather than individual children. This change in research question simplifies the investigation of structural interventions for early childhood education centers, such as hours of operation, full-day versus part-day programs, 1- versus 2-year programs in which children enter either at age 3 or age 4, variation in the length of the school year (traditional 9-month versus full-year), variation in the size of preschool classrooms (paralleling work on class size done at the elementary level), variation in the training or formal education of early childhood education teachers, variation in the socioeconomic composition of classes (mixed versus homogeneous), or variation in the age mixture within classrooms (mixed versus homogeneous classes).

We suggest that changing the research question also leads to a change in research design, such that randomization of groups often is more appropriate than randomization of individuals. The group-randomization strategy is applicable to studies of the comparative effectiveness of such curricular interventions as models for enhancing early literacy (e.g., Breakthrough to Literacy, Let's Begin With the Letter People, the Waterford Early Reading Program, Ladders to Literacy, Partners for Literacy, Core Knowledge, Children's Literacy Initiative, Curiosity Corner, the Creative Curriculum, the High/Scope curriculum, and many others). It can be used to test preschool

programs that have a strong focus on literacy (e.g., those emphasizing dialogic reading and phonemic awareness), alternative delivery models (e.g., curricula that emphasize computer- or Internet-based delivery or those that emphasize professional development of teachers); preschool curricula that stress parenting education programs (e.g., Parents as Teachers or Home Instruction for Parents and Preschool Youngsters); or formal, center-based care versus informal care provided by family members or friends. If the curricular innovations include exemplary model programs, such as the Abecedarian Project or the Perry Preschool program, our proposed strategy could also provide replications essential to determining the consistency and robustness of favorable outcomes across a variety of sites.

At the initial stage, it would make sense to compare program models or program components that contrast strongly. Hence, it would make more sense to compare the relative effectiveness of human- versus computer-based instruction than to compare two alternative computer-based instructional routines.

RATIONALE FOR FEDERAL PRESCHOOL AND CHILD CARE PROGRAMS

When standardized tests began to be used in our schools, it soon became obvious that, on average, children from low-income and minority families consistently did worse on the tests even in the earliest grades (West, Denton, and Germino-Hausken 2000). Public concern about this disparity in performance, coupled with strong desires for improving the ability of underprivileged children to achieve in the adult world, led policy makers to define one of the nation's most important and visible social policy goals: improving the early cognitive and social development of children from low-income families. Because the relatively poor cognitive performance of children from low-income families is evident at the point at which children enter school (Puma et al. 1997), substantial attention has been focused on the effectiveness of preschool programs. Since the beginning of the War on Poverty and the establishment of the Head Start and Title I programs in 1965, early childhood education has been seen as a potentially powerful tool for accomplishing the goal of minimizing cognitive disparities at school entry along ethnic and socioeconomic lines.

Although federal interest in early childhood education had its origins in improving the cognitive and social development of children from low-income families, federal involvement in child care stemmed from an interest in providing support services for low-income mothers seeking to move off the welfare rolls. Thus, federal programs such as Aid to Families With

Dependent Children (AFDC) Child Care have traditionally provided child care as a work-related service and have had no specific intention of improving child literacy skills. However, the ongoing federal push to help children from low-income families reach school on a par with their more advantaged peers has led federal child care programs to develop a dual emphasis on child cognitive development and on the provision of safe, nonparental, work-related care. For example, the Child Care and Development Block Grant provides funds for child care, before- and after-school care, and the development of early childhood education programs. Moreover, welfare-to-work demonstration programs have begun to incorporate systematically a two-generation approach; the intent is to improve child cognitive outcomes as well as move families off the welfare rolls (e.g., Huston et al. 2001; Quint, Bos, and Polit 1997).

Good reasons exist for the optimistic view that early childhood education and child care programs can help children from low-income families. Several randomized experiments evaluating demonstration programs have produced evidence that high-quality, intensive, center-based programs can produce both short- and long-term benefits for children (e.g., Barnett 1995; Bowman, Donovan, and Burns 2001; Karoly et al. 1998; Wasik and Karweit 1994). These findings have been generated, for the most part, from studies of relatively small, carefully designed, model programs that provide intensive, center-based services (e.g., the Perry Preschool Project, the Infant and Health Development Project, and the Abecedarian Project). Unfortunately, none of the model preschool programs evaluated in those studies have been replicated. Consequently, we do not know whether their impressive outcomes can be achieved consistently across a variety of sites and populations. Still, findings from small-scale model early childhood education programs have been used to bolster political and financial support for a wide range of federally funded early childhood interventions.

PAST EVALUATIONS

Prior to the past few years, efforts to evaluate and improve early childhood education and child care programs for low-income children were not very successful, as a result of weak programs and poor evaluation designs. A haphazard approach characterized program development at the national level. Federal officials operate within short-term political realities because Congress and other constituents are not eager to make long-term plans; to fund studies that take years to complete; and to engage in slow, gradual, incremental improvements to educational programs that bear fruit long after

they have left for other positions. Ideas from policy makers often are tied to programmatic approaches—some stale and some innovative—that rarely have been rigorously evaluated. Frequent political shifts impede development of institutional memory among agency officials; instead, officials often have moved from idea to idea with little understanding of what has (or has not) worked well in the past. Demonstration programs appeared, were evaluated, and disappeared; they had no connection to what was previously learned and no discernable effect on what happens in the future. Until recently, few long-term, systematic investigations examined which approach to solving a given problem might work best.

In addition to these problems with designing improved early childhood interventions, much of the research and evaluation on early childhood programs has been deeply flawed, as shown in the other articles in this volume. The research overly relied on weak, quasi-experimental approaches, such as posttest-only and pretest-posttest designs; correlational analyses of observational studies; and case studies (Cook and Payne 2002; Whitehurst 2003). Although randomized experiments are the best way to determine the effectiveness of an intervention because they most strongly support conclusions about the causal connection between treatment and outcomes (Meinert 1986), it has been common in reports from studies of early childhood programs to find claims of causality (e.g., Program X had the following effects) in the absence of a randomly assigned control group, or even a comparison group of any sort. Until recently, even when randomized experiments were used in studies of federal early childhood programs, they typically were conducted as stand-alone efforts, one large study at a time, with limited interconnections and, hence, little cumulative learning.

Head Start, the nation's flagship preschool program, provides the best example. Head Start has spent more than three decades studying the effectiveness of demonstration programs, sometimes using experiments. In the 1970s, Head Start funded a planned variation study to test the relative effectiveness of various approaches to providing Head Start services (Weisberg 1974) and initiated evaluations of the Home Start program (High/Scope Educational Research Foundation and Abt Associates 1976), the Head Start Parent-Child Centers (Holmes et al. 1973), and the Child and Family Resource Program (Travers, Nauta, and Irwin 1982).

After a lull in Head Start research in the 1980s, a resurgence of Head Start demonstration and evaluation activities took place in the 1990s, including randomized experimental studies (at the child level) of the Comprehensive Child Development Program (CCDP; Goodson et al. 2000), the Head Start Family Service Centers (Swartz, Bernstein, and Levin 1998), the Head Start Transition Study (Ramey et al. 2000), and the Head Start Family Child Care

Demonstration (Faddis and Ahrens-Gray 2000). These random assignment evaluations were independent and involved little overlap in areas of investigation, measurement, and design. In the late 1990s and early 2000s, Head Start continued to study the effectiveness of demonstration programs by funding an evaluation of the Early Head Start program that randomized participants at the child level (Love et al. 2002). In addition, the federal government has funded an ongoing national survey of the progress of Head Start children (Zill et al. 2001) and, upon the insistence of the General Accounting Office (GAO) and Congress, a random assignment evaluation (at the child level) of the Head Start program itself (U.S. Department of Health and Human Services [HHS] 2005).

Although the investment that Head Start has made in research and evaluation is commendable, much of the federal research cited above involved stand-alone efforts rather than building blocks in a larger plan of program improvement. For example, the CCDP, funded as a Head Start demonstration program in the 1990s, was a virtual replication of the Child and Family Resource Program, which was funded by Head Start a decade earlier and found to be ineffective. Moreover, little evidence indicates that the wealth of research conducted by Head Start has been used to improve outcomes for children. Only recently have data from the Head Start Family and Child Experiences Survey (FACES) study been used to prod Head Start into emphasizing the importance of literacy-related activities.

CURRENT EVALUATIONS

Although weak research methods have been used in studies of many early childhood programs, the evidence cited above shows that research has started to move in the right direction. In particular, federal early childhood education has seen a substantial increase in the number of high-quality research studies (Greenberg and Shroder 1997). The studies have been methodologically rigorous and often based on random assignment of individual children. Federal agencies, at the insistence of Congress, have called for randomized studies to evaluate the effectiveness of their programs, and such studies are typically accorded greater weight in the decision-making process (St.Pierre and Puma 2000).

In recent years, researchers have accelerated the use of rigorous evaluation methods and the design of coordinated investigations of approaches for improving early childhood education. The most visible and active organization in the move toward increased research rigor is the Institute of Education Sciences (IES) in the U.S. Department of Education. Profoundly influenced

by the No Child Left Behind Act of 2001 and its central principle that federal funds should support educational activities "backed by scientifically-based research" (Coalition for Evidence-Based Policy 2002, 1), the IES has launched a multipronged attack on weak research. First, it has funded several large-scale, high-quality randomized experiments, designed to test the relative effectiveness of various interventions. Second, it is refocusing the Department of Education's research laboratories and centers so that they are producers and disseminators of high-quality, experimentally based research. In particular, the National Center on Early Childhood Education and Development will be charged with testing the effectiveness of different models of professional development for early childhood teachers and comparing the effectiveness of different models for coordinating early childhood programs. And third, it seeks to address the lack of professional research talent by supporting the "training of postdoctoral fellows interested in conducting applied educational research, and to produce a cadre of education researchers willing and able to conduct a new generation of methodologically rigorous and educationally relevant scientific research" (U.S. Department of Education n.d.-b).

As just a few examples of the research currently being undertaken, the IES is sponsoring the following studies:

- *A series of grants for randomized experimental studies of the relative effectiveness of various preschool curricula.* Thirteen such grants were awarded in 2002 and 2003; 10 are randomly assigning intact classrooms to implement a specific new curriculum or to continue with the current curriculum, 2 are randomly assigning schools, and 1 is randomly assigning individual children.
- A randomized experimental study of the relative effectiveness of four different family literacy curricula aimed at promoting literacy and other school readiness outcomes for children. One hundred twenty Even Start projects were recruited and randomly assigned either to implement one of the four family literacy interventions (intervention group) or to continue the current Even Start program (control group).
- *An evaluation of the Reading First program.* This study uses a high-quality regression-discontinuity design involving 250 schools from 20 school districts. One hundred twenty-five schools are participating in Reading First because the scores assigned to their funding proposals were higher than a specified cutoff, and 125 schools did not receive Reading First funds because their proposals fell below the cutoff score.
- *An evaluation of the Early Reading First program.* This study uses a high-quality regression-discontinuity design in which the treatment group consists of children attending preschool in 28 Early Reading First sites and the comparison group consists of children attending preschools in 40 sites that applied for, but did not receive, Early Reading First funding.

- *An evaluation of reading comprehension programs.* This is an evaluation of four different reading comprehension programs with direct instruction in science or social studies content areas. A sample of elementary schools will be recruited and each school will be randomly assigned to implement one of the programs.
- *Evaluation of academic instruction for after-school programs.* Two research-based school curricula have been adapted for use in after-school settings. The curricula are being tested in two evaluations (one for reading and one for math) with 25 centers in each study. After-school program participants at each center were randomly assigned to implement the intensive academic curriculum or to the academic activities (usually homework help) that the center usually provides.
- *Evaluation of remedial reading programs.* This is an evaluation of intensive remedial reading programs for third and fifth graders who have not yet acquired the reading skills necessary to succeed in school.
- *Evaluation of math curricula.* This is an evaluation of the relative effectiveness of five different elementary math curricula. A sample of 100 elementary schools will be recruited and randomly assigned to implement one of the curricula.

For additional information see U.S. Department of Education (n.d.-a, n.d.-c).

The U.S. Department of Health and Human Services (HHS) also has initiated several high-quality studies of early childhood programs. Some of these include the following:

- *Head Start Impact Evaluation.* This evaluation seeks to identify the impacts of Head Start on participating children. Six thousand Head Start–eligible children from 75 grantees have been randomly assigned to participate in the program or to a control group that does not participate (HHS, Administration for Children and Families, Office of Planning, Research, and Evaluation n.d.-b).
- *Early Head Start evaluation.* This evaluation seeks to identify the impacts of Early Head Start on children. Three thousand children from 17 Early Head Start projects were randomly assigned to participate in the program or to a control group (HHS, Administration for Children and Families, Office of Planning, Research, and Evaluation n.d.-a).
- *Child Care subsidy evaluations.* This is an evaluation of three different curricula for improving the language and literacy teaching skills of low-skilled staff, then conducting an evaluation of the language and literacy skills of children in subsidized day care centers that receive CCDF subsidies. The evaluation, being conducted in Dade County, Florida, includes random assignment of 162 child care centers serving 2,000 children to the various language and literacy curricula. A second child care subsidy evaluation is being conducted in Massachusetts, where 350 family day care providers serving children younger than 30 months of age have been randomly assigned to be taught to implement either an enhanced literacy-based curriculum or a standard developmentally appropriate curriculum (HHS, Administration for Children and Families, Child Care Bureau n.d.).

The studies listed above and the other related activities being undertaken show that the early childhood research community has moved into a new era, one in which research rigor is replacing research mediocrity and one in which coordination and systematic knowledge building is replacing fragmentation and haphazard approaches.

REDEFINING THE RESEARCH QUESTION

In the 1980s and 1990s, the best-designed evaluations were randomized experiments that attempted to estimate the effectiveness of early childhood education programs. The research question addressed was, "Does Program X work?" The outcomes of children randomly assigned to attend Program X were compared with the outcomes of children randomly assigned to a control group of children who were not attending that program. This strategy is especially appropriate when it is possible to maintain the treatment integrity of the control group, that is, when it is possible to ensure that children assigned to the control group do not participate in any preschool programs. For example, ensuring control group integrity was possible in the Perry Preschool study (Schweinhart, Barnes, and Weikart 1993) because few, if any, competing preschool programs for disadvantaged children were available in Ypsilanti, Michigan (the site of the program), at the time of the study.

No "no-treatment" groups. While randomized experiments were becoming more prevalent, preschool programs for disadvantaged children were instituted in almost every American community, undermining in part the interpretability of research, even randomized studies, on early childhood education. Today, some sort of early childhood service is available (if not always used) to almost all low-income families, for at least some period of time in a preschooler's life. Thus, a study in which children are randomly assigned to be in Even Start or a control group, Reading First or a control group, or Head Start or a control group ends up comparing two different types of interventions. One intervention is the program of interest; the other is the assortment of early childhood services obtained by the control group. Under current conditions, randomly assigned control groups of individual children are *not* subject to "no-treatment" but to a mixture of other treatments.

Furthermore, mothers who volunteer to participate in a random assignment evaluation of, say, Head Start, are almost always looking for some form of care for their child. The implication for experimental studies of early childhood education programs is that mothers of children who are assigned to the

TABLE 1: Intervention and Control Children Receiving Early Childhood Education in Evaluations of Three Early Childhood Education Programs

Program	*Intervention Children (%)*	*Control Children (%)*
Even Start	72	33
Early Head Start	43	27
Comprehensive Child Development Program	61 (age 4) 51 (age 3) 48 (age 2)	45 (age 4) 29 (age 3) 22 (age 2)

control group are likely to try to enroll their children in a program that provides services similar to those of Head Start, such as Title I preschool or a state-sponsored preschool. Even when children assigned to control groups are enrolled in center-based child care, they often receive some educational services. In many communities, all of these programs are coordinated and may even share physical space. As a result, low-income families can easily access most of them.

Data supporting this contention come from recent experimental studies of three federal early childhood programs: Even Start (St.Pierre et al. 2003), Early Head Start (Love et al. 2002), and CCDP (Goodson et al. 2000). In each of the three studies, data were collected on the extent to which children in the intervention group and in the control group participated in a center-based early childhood education program (Table 1). In each study, a higher percentage of children in the intervention group than in the control group received education services, but it is evident that a substantial percentage of children in each study's control group did not receive "no early childhood education." Instead, they were enrolled in a variety of early childhood services. For example, in the Even Start study, parents reported that control children were enrolled in Head Start, state preschool programs, Title I preschool, and early special education programs. In addition to the children who participated in these formal preschool settings, other control-group children might have been in center- or home-based child care programs.

Because children assigned to control groups so often find alternative, competing early childhood services, it is not surprising that the evaluations cited here did not find substantial program effects. The studies of Even Start (St.Pierre et al. 2003) and CCDP (Goodson et al. 2000) found no impact. Although the Early Head Start study (Love et al. 2002) found some statistically significant program effects, they were small, generally about one-tenth of a standard deviation in magnitude, and concentrated in certain subgroups. Similarly, first-year findings from the Head Start impact study reveal small,

statistically significant, and positive program impacts on some measures (HHS 2005).

An environment saturated with competing alternative programs also affects the treatment integrity of intervention groups. Parents whose children have been assigned to an intervention may move away, decide some other program would be better or more convenient for their child, decide that keeping the child at home would be better, and so on. Because no ethical or legal way exists to force a child to participate consistently in any early education program, an intervention group cannot maintain absolute integrity. Maintaining that integrity can be especially difficult when many alternatives are available in the environment.

Having a saturated preschool environment means that findings from randomized experiments using mixed-treatment control groups or intervention groups with many dropouts, as described above, do a relatively poor job of answering the research question, "Is Program X effective?" Instead, such experiments answer the question, "Is Program X more effective than a mixture of no program and existing competitive programs?" The answer may be of interest to some stakeholders, but it does not contain critical information that can be used to improve program effectiveness. The answer does not tell us which early childhood education intervention works best, which curricular approach helps children most, or whether more intensive interventions work better than less intensive ones. Instead, the result is a series of studies in which Head Start, Early Head Start, Even Start, CCDP, Title I, Reading First, and so on, are compared with mixed-treatment control groups; they provide no systematic way of determining which approach is most effective or which program is the best use of taxpayer dollars.

Comparative assessments. The GAO and the Office of Management and Budget (OMB) each have called for comparative program assessments (OMB 2002). GAO (2000b) has studied the overlap among federal early childhood programs; compared the Head Start and Even Start programs (GAO 2002); and tried to assess the relative effectiveness of early childhood education programs including the Child Care and Development Fund, Head Start, the Social Services Block Grant, and Title I (GAO 2000a). The comparisons were not entirely successful, because they had to be made indirectly.

Because the GAO and the OMB reflect the interests of Congress and the executive branch, respectively, we infer that the agencies are starting to ask more than "Does Program X work?" Instead, they want to know the answers to questions such as, "Which program works better?" and "Which expenditure of federal dollars is most helpful to low-income children?" The idea of systematically investigating alternative approaches for the improvement of

early childhood education leads us to change research questions and associated research designs. Hence, instead of asking whether Program X works when compared with anything except Program X, interest is growing in trying to find out whether Program X, Program Y, or Program Z works best (Boruch and Foley 2001; Boruch et al. 2003).

Policy makers, researchers, and program operators are coming to understand that studies comparing the relative effectiveness of different intervention approaches are often more useful than single-program studies in building knowledge about which strategies for early childhood education work best. For example, the advisory committee formed to help design the recently implemented Head Start Impact Study recommended the comparative approach. Although the committee was charged with designing a Head Start versus a no–Head Start study (and did so), the group recognized that more knowledge about improving the program would be generated by studies in which regular Head Start would be compared with an enhanced or altered Head Start (e.g., 1 versus 2 years of service or an enhanced focus on literacy services). The group suggested "a sequence of these studies with randomization at the site level so that new information about various program options could continuously be used to reshape the core Head Start program" (Advisory Committee on Head Start Research and Evaluation 1999, 100). A similar recommendation came from a design conference on Communities That Care, sponsored by the Office for Substance Abuse Prevention: "A rigorous evaluation of a comprehensive community intervention requires an experimental design whereby communities are randomly assigned to experimental and control conditions" (Peterson, Hawkins, and Catalano 1992, 582). This perspective also is expressed in the National Academy of Sciences' recommendations to the Centers for Disease Control and Prevention on how to evaluate AIDS education programs (Coyle, Boruch, and Turner 1989), and Rossi (1999) recommended this approach in the evaluation of community development programs, as did Hamilton and Rossi (2002) for program development of food assistance programs.

RANDOMIZING GROUPS VERSUS INDIVIDUALS

Given the present ecology of preschool programs, we believe that the "What works better?" question can be better addressed by using research designs in which groups of individuals, such as classrooms, schools, or entire communities, are the units of randomization, instead of individual children. The comparative advantages and disadvantages of individual- and group-randomization evaluation designs are summarized in Table 2 and discussed in the sections below.

TABLE 2: Summary of Design Features Comparing Randomization of Individuals With Randomization of Groups

Design Feature	*Random Assignment of Individuals*	*Random Assignment of Groups*
Primary research question	Does the program work? Determines program effectiveness compared with no-program condition	Which program or program component works better? Determines which of two or more programs (or program components) is more effective
Conditions for optimal use		
Intervention delivery mode	Administered to individuals, families, or households	Administered to groups: classes, schools, neighborhoods, or communities
Program originality	Demonstration of innovative program	Improvement of ongoing program
Comparison	No-treatment condition	Alternative programs
Program ecology	Possibility of competing programs contaminating controls	No no-treatment control
Resistance to randomization		
By program staff	Possibly high due to concern about denial of services	Minimal resistance
By program participants	Possibly high in control group due to disappointment over getting no intervention	Minimal resistance
Postrandomization issues		
Crossovers	More likely: Control group may need intervention-like services.	Less likely: All groups receive desirable services.
Program dropouts	Irrelevant: Dropout is related to program acceptability.	Irrelevant. Dropout is related to program acceptability.
Study attrition	Little difference: Attrition is a function of effort devoted to retaining individuals in study.	Little difference. Attrition is function of effort devoted to retaining individuals in study. Group attrition is rare.
Analysis complexity	Relatively simple intervention-to-control comparisons are used.	Complex multilevel models are needed.
Sample size of individuals needed for given power	Smaller	Larger
Costs	Lower due to smaller sample size	Higher due to larger sample size

Implementing individual randomization designs. Experiments that randomize individuals could be used to answer the "What works better?" question if it were possible to identify control-group conditions that could be maintained as "no treatment." If that could be done, one design approach would be to run several experiments in parallel, each one comparing a preschool program (or curriculum) with "no treatment." If each study worked with the same population of children, then results from the various experiments could be compared and conclusions drawn about which of the interventions worked best. But as argued above, the prevalence of competing preschool services means that that approach is not possible in today's society.

Next, we might try to answer the "What works better?" question by randomly assigning children to participate in different preschool programs. However, randomly assigning children to various early childhood interventions (e.g., Head Start as opposed to Title I preschool) could only be done in areas in which programs exist and transportation to each program is available. (This limitation does not apply to group-randomized designs because the intervention is assigned to existing projects, sites, or classrooms, not individuals.)

In some cases, legal constraints limit the random assignment of individual children. For example, most federal early childhood programs use family income as an eligibility criterion, so it is not possible to enlarge the pool of eligible families by going above the prescribed income limits. Other programs have rules stating that a percentage of children must have certain characteristics; Head Start, for example, mandates that at least 10% of children served must have a disability.

Objections to randomization of individuals. Random assignment of individuals to intervention or control groups is often met by a host of objections from program implementers (Gueron 2002; St.Pierre 2001). An underlying and often unvoiced objection to random assignment of individuals is that program staff typically assume that early childhood education programs are effective, especially their own programs. Concerns about whether scarce programmatic resources are being spent in the most appropriate manner appear to them to be baseless. Perhaps their assumptions are for the better, because scientific skepticism may undermine the ability of program operators to do their best work.

Program operators also resist randomization of individuals because they view it as unethical or believe it means denying services to eligible children. The strong arguments in favor of random assignment of children are often to no avail. Under normal circumstances, program staff have control over which eligible children are served by an early childhood program, and they believe that they can select participants who can derive the greatest benefits from

their program. The crux of resistance to random assignment of individuals is that an experiment changes the mechanism for selecting exactly which children are to be served. Randomization deprives program staff of the power to make assignments and makes them feel that children assigned to the control group are being deprived of services. Thus, in spite of the best arguments offered by researchers, program staff cling tenaciously to the idea that random assignment of individual children equates to denial of services.

Whatever mechanisms are at work, it is difficult to persuade program operators to go along enthusiastically with evaluations involving random assignment of individuals. Federal agencies have been reluctant to require that program operators participate in randomized experiments and instead have counted on the expertise of research staff to persuade program staff to participate. (The recent Head Start impact study is a notable exception.) Unfortunately, researchers have limited options for persuasion. They appeal to the best instincts of program staff (by explaining that a random assignment study is the best way of showing whether this intervention works), offer small monetary or in-kind incentives, offer project staff the opportunity to meet with others participating in the study (through annual project meetings), and offer the chance for some free publicity (through published case studies). None of these inducements are substantial, however, and as a result, the percentage of projects that are willing to participate in experimental studies can be low. For example, with the kinds of incentives outlined above, only about 15% of more than 100 projects eligible to participate in a random-assignment evaluation of Even Start agreed to participate (St.Pierre et al. 2003). The ongoing Head Start impact study, however, had virtually perfect agreement to participate from a national sample of projects. The high rate of participation was achieved by organizing strong political support for the study among the HHS, the National Head Start Association, and other Head Start supporters. When program operators were reluctant to participate, HHS made it clear that cooperation in the research was a condition for continued funding.

Hence, voluntary participation strategies can lead to unacceptably low take-up rates and permit strong selection biases to confound outcomes, whereas approaches that involve federal support and, when necessary, federally mandated participation, work much better. The preferred route for enhancing the validity of evaluations is for federal agencies to mandate participation as a condition of receiving federal funds, even if political problems are involved in doing so. Most early childhood programs are popular and wield substantial political influence. An attempt to make funding contingent on participation in research activities is likely to be met by opposition from those projects, the Congress, and others. Still, the issue of mandated participation in experimental

studies needs to be confronted and the merits debated. Given the poor take-up rates that have resulted from having researchers recruit projects to participate in evaluations, federal officials need to make mandated participation a cornerstone of future research efforts.

Implementing randomized group designs. Not only is the random assignment of groups, such as classrooms or schools, conducive to research that will yield more knowledge, this approach obviates many of the operational issues involved in the random assignment of individuals. First, early childhood programs are delivered not to individual children, but to groups of children, such as classes, sites, or projects. Consequently, the appropriate unit of randomization is not individual children, but the groups of children who are the targets of the program.

Second, random assignment of groups instead of individuals is appropriate for answering the "What works better?" question. The Head Start advisory group recommended "randomization at the site level." In this approach, entire sites, projects, or classrooms would be the units of randomization. Thus, the relative effectiveness of Program X versus Program Y might be evaluated by randomly assigning a group of sites to implement X or Y, and children in the sites would be assessed just as though child-level randomization had been done. The focus is not simply on "Does it work?" but on "What works better?"

Changing the research question, as suggested above, has limitations. Comparing two alternative program models does not directly address whether either program works better than no program. If two models are markedly different in effectiveness, then it is reasonable to expect that the more effective program would also be more effective than no program at all; however, the better program may be ineffective, whereas the less effective program actually may have adverse effects. Similarly, two programs that are not differentially effective from each other may be equally effective (or ineffective) when compared with a no-program control group.

Many research studies have successfully involved the random assignment of units other than individuals. Boruch and Foley (2001) cataloged more than 50 different studies, and the Campbell Collaboration (http://www.campbellcollaboration.org/index.html [accessed May 26, 2005]) lists more than 200 "cluster randomized trials," which use communities, schools, classrooms, or other organizations as the unit of allocation in randomized field trials of a wide variety of interventions (e.g., universal free breakfast, crime prevention, smoking cessation, substance abuse avoidance, violence reduction, nutrition education, fertility control, mathematics education, health education for the elderly, and reduced class size). Furthermore, many of the

studies that are being conducted by IES and HHS (cited earlier) involve group randomization, with units of random assignment that include Even Start projects, family day care providers, day care centers, schools, and classrooms within schools.

Finally, the William T. Grant Foundation has launched an initiative to build capacity related to the design and implementation of group randomized trials (William T. Grant Foundation n.d.). To date, this pioneering initiative has included a conference for practicing researchers, a consulting service provided in conjunction with Steve Raudenbush of the University of Michigan and Howard Bloom of Manpower Demonstration Research Corporation (MDRC), and a Web site that provides relevant scientific materials and other resources.

One of the key benefits derived from the group random-assignment studies described above is that the approach avoids many of the pitfalls associated with randomization of individual children or families. In particular, no site is assigned to a no-treatment control group—every site is assigned to an intervention. Consequently, all arguments about ethics and legal constraints are circumvented. Every site gets an intervention, no one is denied service, no parents are upset about their child being assigned to a no-treatment control group, and no program staff are upset about losing control over who participates.

In addition, group randomization provides information on whether the programs being tested are affected by site characteristics. A truly effective program is one that is robust: It should work well in most sites and have little variation by region, degree of urbanization, or demographic composition. Such a program would be a prime candidate for implementation in every preschool center. In contrast, a program that works well only in sites with certain characteristics may not be worthwhile as a template for a national program.

Despite these benefits, difficulties are associated with randomizing sites within ongoing programs. Federal agencies will have to be convinced of the usefulness of the approach and will have to be willing to mandate participation for local grantees. Opposition to these ideas is likely because each federal program has one or more interest groups devoted to maintaining the program status quo and increasing program funding. For example, if Head Start is to serve as a research platform for studying the effectiveness of early childhood interventions, then it will be important to obtain input from, as well as the sponsorship of, the National Head Start Association. Individual Head Start grantees are politically powerful, so even with a federal mandate and support from the National Head Start Association, substantial resistance may develop at the local level. Still, the benefits of group

randomization for local grantees are so obvious that resistance to participation in experimental studies should be significantly less than for studies that call for randomization of children or families.

Sample size issues. Designs involving group randomization often are more expensive than designs requiring individual randomization. In particular, to achieve a given level of statistical power, more sites must be used and more individual children must be involved than in studies in which the individual child is the unit of analysis, multiplying the costs of data collection. A complete discussion of sample size issues is beyond the scope of this article and interested readers should visit the William T. Grant Foundation Web site (http://www.wtgrantfoundation.org) and examine the papers contained there. Of particular relevance is work by Raudenbush (1997) and Bloom (2004).

The sample size required for any given study will differ for designs requiring higher levels of statistical power, a larger number of model programs, larger or smaller numbers of children in each site, or effect criteria with different statistical properties. The designers of a group-randomization experiment would be well advised to examine closely the sensitivity of the results of power calculations to variations in assumptions and pick a design that has enough power for likely variations that will be actually encountered. The analysis strategy for randomized group experiments is more complicated than for individual randomization counterparts. Complications arise because individual children are nested within groups, and Berk (2005, 422) pointed out that "researchers who design group randomized trials do not usually make the group the unit of analysis," an incorrect procedure that leads to violations of statistical assumptions. Still, as long as researchers use the correct unit of analysis, several appropriate statistical models are available (Murray 1998).

Crossovers, program dropouts, and study attrition. It is relatively easy to plan and implement random assignment of either group- or individual-based experiments. Even when successfully achieved, however, randomization is difficult to maintain over time. In particular, crossovers, program dropouts, and study attrition plague many experiments by creating undesirable changes in the composition of program or study participants over time. Crossovers are participants in a control group who obtain intervention (or intervention-like) services. Program dropouts are participants in an intervention group who do not participate or who leave the intervention. Study attrition occurs when participants in an intervention or control group cease participating in a study.

Although crossovers are a threat to the validity of both group and individual randomization experiments, they are especially worrisome in individual

randomized experiments. In the latter, a crossover typically occurs when the mother of a child assigned to a control group finds a way to obtain services for her child that are similar or identical to the services received by children assigned to an intervention group. Crossovers produce control groups that are not no-treatment groups but instead are combinations of no-treatment and participation in competing programs. This problem affects individual-randomized studies more deeply than it does group-randomized studies because the latter include all children in a desired intervention as an integral part of the design.

Program dropouts are individuals assigned to an intervention who do not show up or who participate for very short periods of time. They are not a threat to the validity of a study because dropout is a normal program occurrence and, in fact, is an indicator of program acceptability. For example, data collected from 18 projects participating in the national Even Start evaluation show that 35% of the families randomly assigned to participate in Even Start never appeared on the information system maintained by project staff (St.Pierre et al. 2003). Individual program dropouts occur with equal frequency in individual- and group-randomization studies. Although studies based on group randomization may suffer from the dropout of entire sites, such an occurrence is rare.

Study attrition affects both individual- and group-randomized studies and is a potential threat to the validity of a study. Study attrition means that some individuals do not participate (or cease to participate) in data collection activities and hence are lost to the evaluation. Attrition occurs for many reasons and needs to be distinguished from dropouts from program services. The former signals that a research team is unable to collect data from a family and is a potential threat to the validity of the study, whereas the latter is a normal occurrence in the operation of any early childhood education program and should not prevent a family from providing data. In this discussion, we are concerned only with study attrition. For example, families may move within the same city or general area; without new contact information, it may be impossible to find them. They may move to a different city or state, so that collecting data from them would be impractical. Parents may get a job that makes them less available for data collection, or they simply may refuse to participate in data-collection activities.

Group-randomized designs are less vulnerable to crossover problems than are experiments involving randomization of individuals; both types of studies, however, face serious attrition problems involving missing observations arising from failure to obtain cooperation from families. Attrition is not an insurmountable problem and can be minimized by intensive efforts to track participants over time and to maintain their goodwill. The better

survey organizations have worked out procedures that can reduce attrition to manageable levels. Of course, such procedures are costly, making a well-run experiment with acceptable attrition levels, whether group or individual based, very expensive.

Although dropout is to be expected in the operation of any early childhood education program and is a phenomenon to be studied as part of an evaluation, study attrition can seriously weaken an evaluation. The threat to validity posed by attrition is the likelihood that the families who stop participating in data collection differ in important ways from the families for whom data are obtained. This potential selection bias makes estimation of program effects problematic. Although statistical adjustments can be made to compensate for bias, using them requires considerable technical skill; moreover, those adjustments are not always effective. Indeed, the best strategy is to make every effort to reduce attrition to the point that such adjustments are not necessary.

PROPOSAL FOR AN "EARLY CHILDHOOD EDUCATION SYSTEMATIC IMPROVEMENT EFFORT"

As argued earlier in this article, federal early childhood and child care programs are not as effective as they might be. Past federal early childhood program improvement efforts have proceeded in a haphazard fashion, and although research on early childhood and child care programs is a mixture of good and bad, the difficulties in implementing and maintaining random assignment of children and families mean that even the best studies have serious limitations. Finally, group-randomized designs offer an opportunity to improve the knowledge generated by research on early childhood programs. To improve early childhood education and have positive effects on the cognitive development of children from low-income families, reform of evaluation efforts is needed.

We propose a potentially more productive approach to bettering the quality of early childhood education in this nation: an extended series of group-randomized experiments that compare promising variations of preschool programs. The approach can be used for two types of evaluations. The first type is concerned with evaluating the relative effectiveness of variations within a single program model. A program typically comprises several elements that can be varied in intensity, duration, staffing requirements, and so on. Questions may be raised as to whether a given mixture of program characteristics is better than another one. For example, should an early childhood program be a full- or half-day program? Should the teaching staff be required

to have special qualifications? How much of the curriculum should be devoted to reading readiness? The research question is, "Which is the better (or best) combination of program characteristics within a single program model?"

The second type of evaluation involves assessing the relative effectiveness of different program models. In this approach, distinctly different program models are compared to assess their relative effectiveness. For example, the High/Scope curriculum may be compared with the Montessori model. The research question becomes, "Is Program Model X better than Program Models Y and Z?" Using either approach, the end result will create better knowledge about what works best and an understanding that can guide the formation of more effective preschool programs for our nation.

We call this effort the Early Childhood Education Systematic Improvement Effort (ESIE). Establishing it will not be easy. The ESIE is envisioned as a long-term, carefully planned, systematic series of experiments firmly anchored in a powerful federal agency and designed to significantly improve early education for preschoolers from low-income families. We seek large improvements, because the compromised development of preschoolers from low-income families is a persistent and troubling problem, and we think that substantial room for improvement exists. Recent research shows that despite the best efforts of Head Start, children who participate in this flagship federal program for preschoolers continue to lag significantly behind national norms (Zill et al. 2001).

The ESIE intervention sponsor. Well-developed early childhood infrastructures already exist in HHS and the Education Department, and the agencies currently administer several programs that are appropriate for systematic experimentation. For example, Head Start and Early Head Start are administered by HHS; and Even Start, Reading First, Early Reading First and Title I are administered by the Education Department. Those agencies are therefore prime candidates for the role of intervention sponsor. Funding for experimentation could come both from existing program funds and from new sources. Each of the programs identified above could dedicate a percentage of its resources to a systematic improvement effort. Few objections should arise from program staff, because the funds would continue to be used to provide early childhood services and would support the extra costs of the programs being tested: additional personnel, staff training, and new services and equipment.

Consider Head Start. It has roughly 2,000 grantees that are funded more or less permanently, and each grantee typically comprises several Head Start classes. The program is stable, having been funded since 1965. It enjoys great

political support and is not going to disappear. Head Start has a rich tradition of research-based attempts at program improvement, but they have been scattered and nonsystematic. As noted earlier, the advisory committee for the Head Start Impact Study recommended that a systematic set of studies be undertaken to improve Head Start, using sites as the unit of analysis. Accordingly, the kinds of studies envisioned as part of the ESIE are not foreign to Head Start. Finally, Head Start has enough money to support the programmatic side of an ongoing research program. Funded at more than $6.8 billion in 2004, 1% of those funds, or $68 million, to sponsor the intervention activities of the ESIE is not an excessive burden. Head Start funds would be used only to support intervention and demonstration activities, not research, which would be funded separately (see below). The same holds for Even Start, which in 2000 distributed $200 million in funding to more than 800 local projects, Reading First ($900 million to local projects), Early Reading First ($75 million to local projects), and Title I preschool programs. None of these programs are as large as Head Start, but still they could be involved.

Head Start has additional advantages as a research platform for the ESIE. In 2002, HHS began planning for a Head Start national reporting system (NRS). Under the NRS, all Head Start grantees will collect and report a common set of start-of-year and end-of-year data on participating children. Data will include the following indicators of school readiness: language development; vocabulary; alphabet knowledge; phonological awareness; numeracy; and for English language learners, progress in the acquisition of English language skills. The NRS also provides for the collection of descriptive data on each Head Start program, center, classroom, teacher, and child. This data collection system was implemented for the first time in the 2003–2004 program year; data are being collected annually thereafter. Thus, Head Start offers a special opportunity for investigating the relative effectiveness of modifications designed to improve the program. Interventions can be developed, studies can be designed, projects can be selected and assigned to alternative interventions, and analyses can be conducted, all using data already being collected through the NRS. This capability assumes, of course, that the NRS data are appropriate outcome measures for the studies in question and that the NRS data collection is properly conducted.

HHS, the National Head Start Association, and the Head Start community have shown their commitment to high-quality evaluations, most recently through the Head Start impact study, which called for recruiting 75 Head Start agencies to participate in a study in which individual children were randomly assigned to Head Start or to a control group. All projects selected for the study agreed to participate, something unheard of in evaluations of

service programs. If HHS, the National Head Start Association, and Head Start grantees could collaborate to successfully implement a study involving random assignment of individuals, then they would find it substantially easier and undoubtedly more satisfying to collaborate on group-randomization studies designed to improve the program.

The ESIE research sponsor. Given the cooperation of intervention sponsors such as the entities listed above, the next step would be to find an entity to serve as research sponsor. The research sponsor would fund all the processes necessary to implement the ESIE, including interaction with the intervention sponsor, formation and meetings of a standing advisory committee, development of interventions to study, selection of sites, provision of incentives, and evaluation of the demonstration activities that are funded by the intervention sponsor.

The research sponsor might be an appropriate office within HHS, such as the Office of Planning, Research and Evaluation in the Administration for Children and Families; the Assistant Secretary for Planning and Evaluation; or a research-oriented agency, such as the Institute of Education Sciences in the Department of Education, the National Institute of Child Health and Human Development, or the National Science Foundation. The major function of the research sponsor would be to administer the ESIE research operations with the aid of the Advisory Committee. Because of the complexity of the development and data collection activities, most of the work may have to be accomplished under contracts with research universities or other research organizations. It would be best if the research sponsor were funded through congressional appropriations added to the budget of the entity chosen to assume that role. Funding should be provided for a relatively long period of time—at least an initial 5 years with optional 5-year extensions. It is possible that some of the major private foundations might be persuaded to become joint funders.

Standing Advisory Committee. A panel of ESIE advisors should be assembled and convened on a regular basis. Advisory Committee members should be chosen on the basis of their substantive knowledge and research expertise. Because they would be expected to provide more than pro forma reviews of activities, they should be paid for their participation. The Advisory Committee would be responsible for helping develop a research plan for the ESIE; for suggesting interventions, alternatives, and models to be investigated; for devising a core research approach to be followed by individual studies (e.g., common selection and randomization methods, baseline and outcome measures for individual children, and measures of the intervention); for reviewing and critiquing reports issued from funded

research projects; and for making recommendations for improving early childhood education practice.

Interventions and models. The Advisory Committee would be responsible for developing and setting priorities among variations (i.e., interventions, programmatic alternatives, and curriculum models) to be tested within the ESIE. This ongoing activity should respond to research findings. Some interventions would be simple and require little development work (e.g., investigating whether it is better to enroll children in part-day or full-day programs). Others would require substantial resources to develop the intervention, provide training materials, monitor implementation, and so on (e.g., development of a structured parenting education curriculum as an add-on to existing Head Start services for children). The development of such alternatives would probably best be accomplished through contracts with research universities or other research organizations.

The intervention sponsor, research sponsor, and Advisory Committee will help develop experimental interventions to be investigated through the ESIE. Development activities would be expected to vary according to whether the evaluation is concerned with testing variations on a single model or differences among different models. In the former case, identification of the single program elements to be varied would be a task for those familiar with the program model. In the latter case, selection of alternative program models and the development of implementation protocols for each model should be undertaken by separate teams, each dedicated to laying out what each considers to be the best version of the model in question.

In either case, the primary criterion for testing interventions or single program variations is that there ought to be a reasonable expectation, consistent with child development theory and prior research, that the intervention will make a sizable difference in the development of preschool-age children from low-income families.

Sequence of the investigations. Research under the ESIE should be conducted differently from previous federally funded research in this area. Instead of funding and evaluating one investigation at a time, the ESIE would involve continual innovation and experimentation. Multiple studies would be ongoing at any given time. The emphasis would be on repeated, connected innovation and on action rather than inaction. The approach would be self-critical: When a particular strategy is not working, researchers will take action to try something different. Under the ESIE, no penalty would exist for advocating and implementing innovations that do not work. It will be difficult to find and develop innovative approaches that make a large difference to the development

of children from low-income families, and it may well be that 75% or more of the innovations will be shown to be ineffective. If substantially improving early childhood education were easy, it would have been done already. Thus, the ESIE will be based on a plan that identifies which innovations to try, which approaches to continue and expand, and which approaches to abandon because they have been shown to be ineffective or harmful.

Experimentation on the kinds of innovations listed above could compare just two alternatives. A strong argument, however, can be made for more complex factorial designs that would enable testing the relative effectiveness of two or more innovations and their interactions. For example, half-day versus full-day Head Start could be crossed with two different curriculum approaches to test the effectiveness of each intervention as well as find out whether one of the curriculum approaches is particularly effective in a half- or full-day format. Three or more curricular innovations could be compared in the same experiment. Complex factorial experiments can arguably provide more information on what works best.

The length of the experiments will probably vary, depending on how much time is needed for preparation and on what the early findings are. As a starting point, each experiment might be designed as a 3-year study including a planning year, an intervention and data collection year (when children are in preschool), and a year of analysis and reporting. Experiments using innovations requiring extensive personnel training or recruitment, such as curricular changes, might require an additional year of training. The premise (until proven otherwise) is that no study would conduct follow-up data collection past the end of preschool unless early analyses show that statistically significant and important large effects are observed at the end of the preschool experience. If no early effects are found, then it makes little sense to search for later effects. Literature on the long-term effects of preschool programs exists, but in early childhood applications, such effects have been found only after short-term effects were first demonstrated—never in the absence of short-term effects. Experiments that show promising preschool effects are good candidates for measurement extending through later years of schooling.

Developing and using measures of the fidelity of planned interventions will be a crucial part of any ESIE study. Measurements should be taken at appropriate times prior to each intervention (e.g., during teacher training) and during its implementation. An intervention should be abandoned if the measurements show that it is not being implemented as planned. Thus, some studies may be started but terminated early due to unsuccessful implementation.

Select sites to implement various alternatives. Once a programmatic alternative is ready to be tested, a mechanism for testing it in a sample of sites

will be needed. The intervention sponsor, research sponsor, and Advisory Committee will have to collaborate and decide jointly whether site-level participation in the intervention activities should be voluntary or mandatory. We believe that participation should be a condition for continued receipt of federal funding. Mandatory participation would allow sites to be distributed across the nation (if desired); would enhance the generalizability of findings; and, over time, would help build a culture favoring ongoing program improvement. In contrast, voluntary participation might permit selection biases, leading to serious flaws that would invalidate ESIE findings.

Independent monitoring and evaluation of each investigation. Each experimental investigation should be conducted by a research team independent of the intervention sponsor and responsible to the research sponsor and the Advisory Committee. The research team will be responsible for designing the experiment, conducting the random assignment and monitoring its integrity, developing measures of treatment fidelity and using them to monitor the implementation of each intervention, taking all possible steps to reduce crossovers and attrition from the study, collecting data on implementation and outcomes, analyzing the results, and preparing research reports. The research sponsor would be responsible for developing and overseeing contracts for the research recommended by the Advisory Committee.

CONCLUSION

For several decades, the federal government has tried to improve the cognitive and social development of children from low-income families, but with limited success. In part, the disappointing results stem from the approach that has been taken to designing and evaluating federal early childhood programs. A better strategy, one that relies on the systematic investigation of alternative interventions through random assignment of sites from existing programs, could help achieve the federal goal of helping children from low-income families someday join the economic mainstream of American society.

REFERENCES

Advisory Committee on Head Start Research and Evaluation. 1999. *Evaluating Head Start: A recommended framework for studying the impact of the Head Start program.* Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth and Families, Head Start Bureau.

Barnett, W. S. 1995. Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children* 5 (3): 25-50.

Berk, R. A. 2005. Randomized experiments as the bronze standard. *Journal of Experimental Criminology* 1:417-33.

Bloom, H. S. 2004. Randomizing groups to evaluate place-based programs. In *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation.

Boruch, R. F., and E. Foley. 2001. The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In *Validity and social experimentation: Donald Campbell's legacy*, vol. 1, ed. L. Bickman. 193-238. Beverly Hills, CA: Sage.

Boruch, R. F., H. May, H. Turner, and J. Lavenberg. 2003. *Estimating the effects of interventions that are deployed in many places: Place randomized trials*. Philadelphia: University of Pennsylvania, Graduate School of Education.

Bowman, B. T., M. S. Donovan, and M. S. Burns, eds. 2001. *Eager to learn: Educating our preschoolers*. Washington, DC: National Academy Press.

Coalition for Evidence-Based Policy. 2002. *Bringing evidence-driven progress to education: A recommended strategy for the U.S. Department of Education*. Washington, DC: The Council for Excellence in Government.

Cook, T. D., and M. R. Payne. 2002. Objecting to the objections to using random assignment in educational research. In *Evidence matters: Randomized trials in education research*, ed. F. Mosteller and R. Boruch, 150-78. Washington, DC: Brookings Institution.

Coyle, S. L., R. F. Boruch, and C. T. Turner, eds. 1989. *Evaluating AIDS prevention programs*. Washington, DC: National Academy of Sciences.

Faddis, B. J., and P. Ahrens-Gray. 2000. *Evaluation of Head Start family child care demonstration: Final report*. Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Portland, OR: RMC Research Corporation.

Goodson, B. D., J. I. Layzer, R. G. St.Pierre, L. S. Bernstein, and M. Lopez. 2000. Effectiveness of a comprehensive five-year family support program on low-income children and their families: Findings from the Comprehensive Child Development Program. *Early Childhood Research Quarterly* 15 (1): 5-39.

Greenberg, D., and M. Shroder. 1997. *The digest of social experiments*. 2nd ed. Washington, DC: Urban Institute.

Gueron, J. M. 2002. The politics of random assignment: Implementing studies and affecting policy. In *Evidence matters: Randomized trials in education research*, ed. F. Mosteller and R. Boruch, 15-49. Washington, DC: Brookings Institution.

Hamilton, W. L., and P. H. Rossi. 2002. *Effects of food assistance and nutrition programs on nutrition and health*. Vol. 1, *Research design*. Report for the U.S. Department of Agriculture, Food and Nutrition Service. Cambridge, MA: Abt Associates.

High/Scope Educational Research Foundation, and Abt Associates. 1976. *Home Start evaluation study. Final report: Findings and implications*. Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Ypsilanti, MI: High/Scope Educational Research Foundation.

Holmes, M., D. Holmes, D. Greenspan, and D. Tapper. 1973. *The impact of the Head Start parent-child centers on children: Final report*. Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Washington, DC: Center for Community Research.

Huston, A. C., G. J. Duncan, R. Granger, J. Bos, V. McLoyd, R. Mistry, D. Crosby, C. Gibson, K. Magnuson, J. Romich, and A. Ventura. 2001. Work-based antipoverty programs for parents can enhance the school performance and social behavior of children. *Child Development* 72 (1): 318-36.

Karoly, L. A., P. W. Greenwood, S. S. Everingham, J. Hoube, M. R. Kilburn, C. P. Rydell, M. Sanders, and J. Chiesa. 1998. *Investing in our children*. Santa Monica, CA: RAND.

Love, J. M., E. E. Kisker, C. M. Ross, P. Z. Schochet, J. Brooks-Gunn, D. Paulsell, K. Boller, J. Constantine, C. Vogel, A. Fuligni, and C. Brady-Smith. 2002. *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start. Executive summary*. Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Washington, DC: Mathematica Policy Research.

Meinert, C. L. 1986. *Clinical trials: Design, conduct, and analysis*. New York: Oxford University Press.

Murray, D. M. 1998. *Design and analysis of group randomized experiments*. New York: Oxford University Press.

Peterson, P. L., J. D. Hawkins, and R. F. Catalano. 1992. Evaluating comprehensive community drug risk reduction interventions. *Evaluation Review* 16 (6): 579-602.

Puma, M. J., N. Karweit, C. Price, A. Ricciuti, W. Thompson, and M. Vaden-Kiernan. 1997. *Prospects: Final report on student outcomes*. Report for the U.S. Department of Education, Planning and Evaluation Service. Bethesda, MD: Abt Associates.

Office of Management and Budget. 2002. *Managing for results: Budget and performance integration: Program performance assessment*. Washington, DC: Office of Management and Budget.

Quint, J. C., J. M. Bos, and D. Polit. 1997. *New Chance: Final report on a comprehensive program for disadvantaged young mothers and their children*. New York: Manpower Demonstration Research Corporation.

Ramey, S., C. Ramey, M. Phillips, R. Lanzi, C. Brezansek, C. Katholi, and S. Snyder. 2000. *Head Start children's entry into public school: A report on the National Head Start/Public Early Childhood Transition Study*. Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Birmingham, AL: Civitan International Research Center.

Raudenbush, S. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods* 2 (2): 173-85.

Rossi, P. H. 1999. Evaluating community development programs: Problems and prospects. In *Urban problems and community development*, ed. R. Ferguson and W. Dickens. Washington, DC: Brookings Institution.

Schweinhart, L. J., H. V. Barnes, and D. P. Weikart. 1993. *Significant benefits: The High/Scope Perry Preschool study through age 27*. Monograph 10. Ypsilanti, MI: High/Scope Educational Research Foundation.

St.Pierre, R. G. 2001. Random assignment: Implementation in complex field settings. In *International encyclopedia of the social and behavioral sciences*. Vol. 19, *Logic of inquiry and research design*, ed. N. Smelser and P. Baltes (series eds.) and T. Cook and C. Ragin (vol. eds.), 12731-34. Oxford, UK: Pergamon.

St.Pierre, R. G., and M. J. Puma. 2000. Toward the dream of the experimenting society. In *Validity and social experimentation: Donald Campbell's legacy*, ed. L. Bickman, 169-92. Beverly Hills, CA: Sage.

St.Pierre, R. G., A. Ricciuti, F. Tao, C. Creps, J. Swartz, and T. Rimdzius. 2003. *Third national Even Start evaluation: Program impacts and implications for improvement*. Report for the U.S. Department of Education, Planning and Evaluation Service. Cambridge, MA: Abt Associates.

Swartz, J. P., L. Bernstein, and M. Levin. 1998. *Evaluation of the Head Start Family Service Center Demonstration projects*. Vol. 1, *Final report from the national evaluation*. Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Cambridge, MA: Abt Associates.

Travers, J., M. Nauta, and N. Irwin. 1982. *The effects of a social program: Final report of the Child and Family Resource Program's infant-toddler component.* Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Cambridge, MA: Abt Associates.

U.S. Department of Education. n.d.-a. National Center for Education Evaluation and Regional Assistance. http://www.ed.gov/about/offices/list/ies/ncee/index.html (accessed May 25, 2005).

———. n.d.-b. National Research and Development Centers—Applicant Information. http://www.ed.gov/programs/edresearch/applicant.html (accessed May 25, 2005).

———. n.d.-c. Study plans of the National Center for Education Evaluation and Regional Assistance. http://www.ed.gov/rschstat/eval/resources/studyplans.html (accessed May 25, 2005).

U.S. Department of Health and Human Services, Administration for Children and Families. 2005. *Head Start impact study: First year findings.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

U.S. Department of Health and Human Services (HHS), Administration for Children and Families, Child Care Bureau. N.d. Main research and data page. http://www.acf.hhs.gov/programs/ccb/research/ (accessed May 25, 2005).

U.S. Department of Health and Human Services (HHS), Administration for Children and Families, Office of Planning, Research, and Evaluation. n.d.-a. Early Head Start research. http://www.acf.hhs.gov/programs/opre/project/tprojectIndex.jsp?topicId=4 (accessed May 25, 2005).

———. n.d.-b. Head Start research. http://www.acf.hhs.gov/programs/opre/project/tprojectIndex.jsp?topicId=6 (accessed May 25, 2005).

U.S. General Accounting Office (GAO). 2000a. *Early childhood programs: Characteristics affect the availability of school readiness information.* GAO/HEHS-00-38. Washington, DC: GAO.

———. 2000b. *Early education and care: Overlap indicates need to assess crosscutting programs.* GAO/HEHS-00-78. Washington, DC: GAO.

———. 2002. *Head Start and Even Start: Greater collaboration needed on adult literacy.* GAO-02-348. Washington, DC: GAO.

Wasik, B. A., and N. L. Karweit. 1994. Off to a good start: Effects of birth to three interventions on early school success. In *Preventing early school failure: Research, policy, and practice*, ed. R. Slavin, N. L. Karweit, and B. A. Wasik, 13-57. Boston, MA: Allyn & Bacon.

Weisberg, H. 1974. *Short-term cognitive effects of Head Start programs: A report on the third year of planned variation, 1971-1972.* Report for the U.S. Department of Health and Human Services, Administration on Children, Youth and Families. Cambridge, MA: Huron Institute.

West, J., K. Denton, and E. Germino-Hausken. 2000. *America's kindergartners: Findings from the Early Childhood Longitudinal Study, kindergarten class of 1998-99, fall 1998.* NCES 2000-070. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Whitehurst, G. J. 2003. The institute of education sciences: New wine, new bottles. Presentation at the annual meeting of the American Educational Research Association, Chicago.

William T. Grant Foundation. n.d. Capacity building for group-randomized studies. http://www.wtgrantfoundation.org/info-url_nocat4030/info-url_nocat.htm?attrib_id=9485 (accessed May 25, 2005).

Zill, N., G. Resnick, K. Kim, R. McKey, C. Clark, S. Pai-Samant, D. Connell, M. Vaden-Kiernan, R. O'Brien, and M. D'Elio. 2001. *Head Start FACES: Longitudinal findings on program performance, third progress report.* Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth and Families.

Robert G. St.Pierre is president of STP Associates, an educational research consulting organization. From 1975 to 2005, he was a vice president and principal associate in the Education and Family Support Area of Abt Associates Inc., where he was principal investigator for educational research, evaluation, and policy analysis projects, conducting randomized experimental studies in diverse areas such as family literacy, case management, compensatory education, curricular interventions, school health education, and child nutrition. He received his BA in mathematics from Northeastern University and his PhD in educational research, measurement, and evaluation from Boston College.

Peter H. Rossi is Stuart A. Rice Professor Emeritus of Sociology and Director Emeritus of the Social and Demographic Research Institute at the University of Massachusetts, Amherst. He has been on the faculties of Harvard University, the University of Chicago, and Johns Hopkins University and was Director, 1960 to 1967 of the National Opinion Research Center at the University of Chicago. He is past-president (1980) of the American Sociological Association and was elected a fellow of the American Academy of Arts and Sciences and of the American Association for the Advancement of Science.