Social Policy Report

Giving Child and Youth Development Knowledge Away

Volume XVIII, Number II 2004

Beyond the Methodological Gold Standards of Behavioral Research: Considerations for Practice and Policy

Robert B. McCall, University of Pittsburgh, Beth L. Green, NPC Research, Inc.

Abstract

Research methods are tools that can be variously applied - depending on the stage of knowledge in a particular area, the type of research question being asked, and the context of the research. The field of *program evaluation*, critical for social policy development, often has not adequately embraced the full range of methodological tools needed to understand and capture the complexity of these issues.

The dominant paradigm, or "gold standard," for program evaluation remains the experimental method. This standard has merit, particularly because experimental research has the capacity to draw conclusions about cause and effect ("internal validity"). This paper identifies the benefits, common problems, and limitations in three characteristics of experimental studies: theory-driven hypotheses; random assignment of subjects to intervention groups; and experimenter-controlled, uniformly-applied interventions.

Research situations are identified in which reliance on the experimental method can lead to inappropriate conclusions. For example, theory-driven hypotheses, if validated, provide a broader base of understanding of an issue and intervention; but some questions should be studied in the absence of theory simply because practice or policy needs the answer. Random assignment can produce cause-and-effect conclusions, but public services are never randomly assigned, and their effectiveness may well depend on participants' motivation or belief in the service (as signaled by their choice to participate). Experimenter-controlled uniform treatment administration insures that we know precisely the nature of the treatment documented to work by the evaluation, but it prohibits tailoring treatment to the individual needs of participants, which is a major "best practice" of service delivery.

Suggestions are offered for ways to incorporate alternative research methods that may emphasize "external validity" (match to real-life circumstances), and complement results derived from experimental research designs on social programs.

The field of program evaluation, and the policy decisions that rest on it, should utilize and value *each* research method relative to its merits, the purpose of the study, the specific questions to be asked, the circumstances under which the research is conducted, and the use to which the results will be put.

Social Policy Report

Editor
Lonnie Sherrod, Ph.D.
sherrod@fordham.edu

Associate Editor
Jeanne Brooks-Gunn. Ph.D.
brooks-gunn@columbia.edu

Director of SRCD Office for Policy and Communications
Mary Ann McCabe, Ph.D.
mmccabe@apa.org

Managing Editor
Bridget Ehart

From the Editor

In this issue Robert McCall and Beth Green examine the methodological tools available for evaluations of programs and policies. It has typically been assumed by funders and policymakers, as well as scholars, that the experiment is the "gold standard." The experimental method is valuable because it allows assignment of causality of any statistically significant outcomes to the program or policy. These authors distinguish internal validity, demonstrated by the experiment, from external validity, which they argue is not always best served by an experiment. They appeal to scholars, policy makers, and practitioners to consider their choice of method in the full context of the research being undertaken. McCall and Green make the case that the experiment, despite its ability to impart causality, may not always be the best method, all else considered. These authors make the important point that the experiment is one of several methodological tools available to the program evaluator. Different methods contribute different kinds of information but one is not necessarily more valuable than another. The value of the method is determined by the questions being asked, the uses to which the research will be put, and so forth. The experimental method, especially in the policy context, has limitations as do other methods; the methods differ in their limitations but none are sufficiently limitation free to constitute "gold."

This is a very important message to impart to the policy and funding community. However, as two of our commentaries demonstrate, not everyone agrees with this argument, including Associate Editor Jeanne Brooks-Gunn, who offers one of the commentaries. We do not usually employ a "point counter-point" style for *Social Policy Report* because policymakers need answers or at least advice, not debate. However, this issue by definition deals with an issue of opinion, that is, choice of research methodology, not with a research finding. Hence, we concluded that the point counter-point approach was merited.

For what it is worth, I do agree with these authors, which is why I commissioned this issue. For years I have been somewhat mystified by our infatuation in evaluation research for the experimental method. True, it is nice to be able to say with the certainty of an experiment that a program or policy caused a certain outcome. Nonetheless, much research in our field of child development is not experimental. We think, for example, that we know something about what makes for effective and ineffective parenting, yet children are not randomly assigned to good and bad parents. Much research on child development is correlational in nature, not experimental. The field of child development has thereby recognized the value of multiple methods. The point of McCall and Green is that program evaluation should do the same. Dr. Cottingham, in her statement, approaches this fact from the opposite viewpoint. She argues that we in fact need more experiments in the field because we have so much non-experimental research. This is a very valid point, but I think these authors make the point that experimental research need not be program evaluations.

At Fordham, we are establishing a Center to address, among other aims, how multiple methods may prove useful in evaluations. Evaluation research is just another form of learning—learning about programs or policies, as well as the children and families that participate in them. In fact, I do not especially like the term "evaluation" because it implies a grading or valuation of what is being studied. Although evaluations can have this tone, it is not an essential aspect. I would like to see evaluations planned that are oriented to learning about the program or policy as well as learning about the phenomena addressed by that program or policy. Pam Morris at MDRC is approaching their evaluations, most of which are experimental, from this perspective; in this way, policy analysis contributes to basic knowledge. Evaluations can contribute to our general knowledge as well as assess a program, and in this way they become not just evaluative, and are thereby less intimidating to the program.

We hope that this issue will serve to orient scholars, practitioners, policy makers and participants to the many useful approaches to program evaluations and to the varied potential contributions of evaluation research. I think all of our authors agree with this point; they only disagree on how best to search for gold!

Lonnie Sherrod, Editor

Beyond the Methodological Gold Standards of Behavioral Research: Considerations for Practice and Policy

Robert B. McCall, University of Pittsburgh
Beth L. Green , NPC Research, Inc.

The gold standard methodology for determining cause and effect in the behavioral sciences is the experimental method, the main components of which are 1) theory-driven hypotheses, 2) random assignment of participants to treatments, 3) experimenter-controlled manipulations uniformly applied to all participants under rigorously controlled conditions, and 4) quantitative measurement and statistical analysis. Although many other methods (e.g., quasi-experimental, qualitative, ethnographic, and mixed-methods approaches) are often used in psychological research, preference and value for the traditional methodological approaches persists.

For example, one leading textbook on program evaluation states that "the gold standard research design for impact assessments is the randomized controlled experiment" (Rossi, 1997, p. 63). Others in the field have suggested that "if intervention research does not rely on random assignment, it does not belong in our scientific journals," and indeed, literature reviews are often confined to randomized studies, ignoring everything else (e.g., Campbell Collaboration Library; Datta, 1994). Moreover, after years of relative disinterest in research, policy makers are now frequently demanding "evidence-based services" (e.g., McCall, Groark, & Nelkin, in press; Weissberg & Kumpfer, 2003); however, researchers have been so skillful at communicating the strengths of the experimental method that many now consider it the *only* way of obtaining valid evidence about social interventions. Indeed, researchers are now often faced with situations in which randomized experimental designs are legislatively mandated, even when the finest methodological scholars in the field would advise against it (Cook, 1998). For example, Ron Haskins, a former staffer for the U. S. House Ways and Means Committee, declared that "unless policy makers have random assignment, the results are definitely suspect… Random assignment is considered real science, real knowledge" (Haskins, 2001). This over-value for experimental methods has been accompanied by an under-value for research relying on alternative methodologies, even when such alternatives may be better suited to answer some key research and policy questions (Schorr, 1999). In short, both policy makers and researchers should recognize that the "evidence" for evidence-based services and policies is not simply and exclusively defined as randomized trials of a uniform treatment, but that "evidence" is much more broadly defined, and has its own set of standards (McCall et al., in press).

The Rationale and Purpose of this Paper

Research methods are tools that can be variously applied depending on the developmental stage of knowledge in a particular area, the type of research question being asked, and the ecological context of the research (e.g., Hedrick, 1994; House, 1994; Sechrest, Babcock, & Smith, 1993; and many others). While this statement is not likely to be debated, we argue that researchers and policy makers do not fully practice it, and experimental methodologies are still too often preferred and other methods ignored or disparaged.

We emphasize that *our purpose is not to denigrate the value of the experimental method*; indeed, randomized designs, conducted under appropriate conditions, have considerable merit, especially for supporting causal inferences. Further, we recognize that in some fields, such as program evaluation in education, the emphasis is inappropriately reversed, and non-traditional methods are *de facto* prevalent at the expense of experimental approaches (Cook, 2002). Thus, our goal is to have all methods appropriately valued for the information they can provide, with accurate and balanced recognition of their assets and limitations. It is the integration of results from a complete set of these complementary approaches that will lead to more comprehensive understanding and more effective services and policies.

> Unless policy makers have random assignment, the results are definitely suspect…

Our focus in this paper is on those researchers, practitioners, and policy makers who, in our opinion, over-emphasize experimental methods. Consequently, we will address the problems and limitations associated with three of the four principal components of the traditional experimental method: (1) theory-driven hypotheses, 2) random assignment of participants to research conditions and (3) experimenter-controlled uniform treatment manipulations. The fourth component, quantitative measurement and statistical analysis has been discussed extensively elsewhere (e.g., Reichardt & Rallis, 1994). For each of these three areas, we discuss the traditional rationale and benefits, the limitations, and suggestions for ways to incorporate alternative methodologies that may improve the state of applied knowledge, especially research on social programs and interventions for children, youth, and families.

SOME LIMITATIONS OF THE EXPERIMENTAL METHOD

Theory-Driven Research

Many funding agencies and journal publication policies demand that research be explicitly guided by theory. Theory, if valid, represents consummate understanding, generality, and breadth of application of cause-and-effect principles; and because it specifies hypotheses about the relations between variables, it restricts the permissible outcomes and reduces capitalization on chance.

Good theory is desirable in applied contexts, even practical (C. Weiss, 1995). There should be a *rationale* for the question, the intervention, the outcomes, the measurement instruments, and especially why the intervention might produce the outcomes to be assessed (i.e., a so-called "theory" of change; e.g., Chen & Rossi,

1987; C. Weiss, 1997); we are not advocating "random" research or evaluation. But the rationale does not necessarily need to be theoretical: Some research should be conducted because society needs to know the answer, regardless of whether *theory* predicts an outcome (McCall & Groark, 2000). Social problems should be detected and parameters described because such problems affect substantial numbers of people, they affect a few people in severe ways, or large amounts of money are spent on a program. Does Head Start produce educational and social benefits for at least some children? Is Head Start too late, and thus Early Head Start is needed to produce desired outcomes in children? Theory building sometimes comes later; that is, theory may be *induced* from research as well as research *deduced* from theory.

Researchers are currently being pressured to contribute both to theory and to societal improvement (McCall & Groark, 2000), that is, they are urged to fall into "Pasteur's quadrant" (Stokes, 1997). Stokes argues that the contributions of research to theory and to practice represent orthogonal dimensions, and the ideal goal is for research to fall into the high-high quadrant, such as the work of Pasteur. That may be ideal, but there is a place for research that "only" contributes to theory (which scientists readily accept) as well as research that "only" contributes to societal benefit (which scientists have often eschewed).

Random Assignment

Random assignment to treatments, typically at the level of the individual, is the key defining attribute of the experimental method, and it is crucial to establishing the internal validity of cause and effect. The "randomized trial" is the gold standard methodology in the health sciences, especially for drug tests, and some psychological researchers have long advocated for the randomized experiment as the primary method for determining the efficacy of social interventions (Campbell, 1969; Cook, 2002). As a result of this successful advocacy, funders of research on social interventions now frequently require evaluation studies to use random assignment to test the impact of new policies and social programs (Conrad & Conrad, 1994; Schorr, 1999; St. Pierre, 1983).

A number of researchers have criticized randomized field experiments, citing ethical issues (Dunford, 1990; Fetterman, 1982), difficulties with implementing and carrying out randomized studies in applied settings (Conner, 1977; Cook & Campbell, 1979), and lack of attention to ethnicity (Sue, 1999). We will not re-visit these criticisms here (see Cook, 2002). Instead, we describe below several common research situations in which randomization at the level of individual participants, even if successfully implemented, may lead to inappropriate conclusions.

Understanding selection into treatment. A major strength of random assignment is the balancing of individual differences (especially those related to self-selection into treatment) across treatment conditions that otherwise may confound a causal conclusion about a given treatment. However, this investment in internal validity may be associated with considerable loss of external validity (e.g., Sue, 1999). Specifically, outside of randomized clinical research trials, people are not randomly assigned to behavioral services that are provided in communities (e.g., children are not randomly assigned to Head Start). Individuals choose or are referred to specific services to meet their needs and participate or not depending on a myriad of factors (McCurdy & Daro, 2001). Further, unlike medical research, double-blind designs and "placebo controls" are not usually available for social experiments; thus, the participants' perception and opinion of the treatment (or of *not* receiving the treatment) may substantially influence the effectiveness of the treatment.

> We are not advocating "random" research or evaluation.

A lack of attention to selection factors can influence the external validity of social research in several ways. For example, people willing to be randomly assigned in a demonstration program may not be the same as those who would choose a particular treatment when it is offered as a community service. This may be especially true among particularly "over-studied" and savvy minority populations, who may be reluctant to volunteer for "a research study" but be quite willing to participate in services that are publicly offered in the community (Cauce, Ryan, & Grove, 1998). They may distrust the research establishment, not want to be "guinea pigs," and not want to risk being assigned to the no-treatment group. In this case, random assignment can lead to "correct" findings about those who are willing to participate in random assignment, but such results may be inappropriately generalized to those who are not willing but who would be consumers of such a service when routinely provided.

Factors related to participants' decisions to be involved in a research study can have other influences on program outcomes as well. For example, the original recruitment protocols for the Comprehensive Child Development Project (CCDP), a federally funded, randomized study of community-based family services, informed potential participants that half of the families would receive program services; the other half would participate in the research and receive fairly substantial stipends in return. Anecdotal evidence suggested that a number of families agreed to participate because they wanted to receive the stipends. When they were assigned to receive the program (with no stipends) they were disappointed and unmotivated to commit to the program's intensive home visiting protocols (McAllister, 1993). One would not expect the program to have the same level of benefits for these families, compared to those who were motivated and interested in receiving services and remained in the study.

Further, suppose divorcing couples were randomly assigned either to having the court decide custody of their children (the "treatment as usual") or to receiving an intervention based on family mediation. It is possible, perhaps even likely, that their children's adjustment will depend on whether the couple would have chosen

Continued on page 6

Beyond Advocacy: Putting History and Research on Research into Debates about the Merits of Social Experiments

Thomas D. Cook, *Northwestern University*

It is a truism that every academic field needs multiple methods because methods are issue-specific and all fields deal with multiple types of issue. For instance, applied developmentalists want to know how coherent individual theories are, how theoretically faithful intervention designs are, who is or is not exposed to an intervention, how good implementation is on the average, why variation in exposure occurs, what effects exposure has, why effects come about, how cost-effective an intervention is, how the results from interventions with similar aims compare to each other, how a study's results fit in with prior evidence on the same topic, and how a theory or intervention should be valued in light of all that is known (Cook, Shadish & Leviton, 1985). At some time or another, each of these questions is important. However, many of them are not causal and so do not require experiments.

The knowledge needs above are listed in a logical sequence. One might proceed by using the best available method for answering the earliest question and, when this is done, then go on to the next issue in the sequence, and so on. A multi-decade research program would result. As rational as this would seem to be, it does not reflect how science progresses. Rather, researchers try to bundle multiple questions into a single study, answering the highest priority one with the best method available and answering others with whatever lesser methods can be fit into the constraints created by the study context and the methods already chosen. In many social sciences, the past priority was on external validity. This led to placing most design weight on the random selection of households and on then measuring many potential cause, effect, moderator and mediator variables that could be subject to descriptive, causal or explanatory analysis. In this schema, internal validity was assigned a secondary status and causal knowledge was pursued via measurement and statistical modeling rather than manipulation and random assignment.

Unfortunately, in their current form statistical adjustments like instrumental variables, propensity scores or simple regression almost certainly produce more biased causal inferences than random assignment. Glazerman, Levy & Meyers (2003) have summarized 12 studies where the effect size found in a randomized experiment was compared to the effect sizes produced by non-experimental analyses that retained the treatment group but used other data sources to replace the original random control group with a non-equivalent one. Whatever the statistical adjustment method used, the effect sizes were rarely similar across the experimental and non-experimental analyses. If one assumes from theory that the experimental estimates are less biased, this suggests that our current repertoire of statistical adjustments is inadequate, though this conclusion also requires us to assume that random assignment was perfectly implemented in the studies that Glazerman et al summarize. In any event, prior faith in statistical adjustments of all kinds has been undermined by this kind of research on research, and bundling research questions within surveys of parents, children or teachers is now widely seen as counterproductive when a high quality causal inference is at stake.

The Glazerman et al findings also imply an attack on quasi-experimental alternatives to the experiment that are based on manipulating the possible cause rather than merely measuring it. This is because the designs abstracted from the surveys compare two or more non-equivalent groups on an outcome, sometimes when there are pretest measures of the outcome and sometimes a priori matching. Also speaking against quasi-experimental alternatives are some brute empirical findings from meta-analytic work. Distributions of effect sizes between experiments and quasi-experiments conducted on the same topic reveal no differences in the means but considerable differences in the variances, with the experiments achieving greater consistency in their causal estimates (Lipsey & Wilson, 1993; Glazerman et al, 2003). This difference in precision has several possible interpretations, but it suggests a novel justification for random assignment in research areas where a large stock of experimental evidence does not already exist. In this circumstance, it is difficult to know whether any single quasi-experiment has under- or overestimated the true causal relationship, breeding considerable uncertainty about the result. In applied developmental science, it seems rare to me to find many experiments on a topic. So it seems dangerous at this time to promote quasi-experiments as alternatives to experiments. However, I would be more comfortable in this if past research had included the more powerful quasi-experimental designs that Shadish, Cook & Campbell (2002) advocate rather than the primitive designs that were included and that Shadish et al reject or are cool about.

If experiments are to be preferred, they should not be black box ones that seek to retain experimental purity at the cost of obscuring the context in which a causal inference is inevitably embedded (Cook, 2000). Unbiased causal inference is possible without standardized interventions and without measures of implementation or other aspects of program theory. But these features are nonetheless crucial. They increase both the precision of causal estimates, the construct validity of the manipulated cause, and the explanation of why molar effects did or did not occur. Also, external validity is enhanced if a study samples (and measures) heterogeneous populations. All this suggests that the basic experimental structure should be complemented by the measurement and analysis of data about people, attrition processes, treatment implementation, and other theoretical influences. However, bundling many questions into an experiment increases measurement burdens, expense and obtrusiveness. It also entails moving beyond the basic intent to treat analyses in order to conduct internal analyses whose results are endangered by the very selection bias that random assignment is designed to avoid. So bundling questions into an experiment has to be done in full knowledge that some questions cannot be answered as well as others. There are no free lunches.

McCall and Green (2004) are correct that the art of implementing field experiments is far from perfect. But consider that survey research is one of the most successful social science tools of our time. It is based on two separate theories. The first is an elegant, abstract sampling theory about how cases are selected. The second is a mundane and diffuse theory, constructed over 60 years, that

specifies better ways to implement surveys through generalizing past research on: (1) what percent of non-responses significantly biases the representativeness of nation-level results, (2) how data should be collected with respect to question wording, response formats, interviewer race effects, face to face versus telephone surveys, and retaining respondents in longitudinal surveys, and (3) how data should be analyzed with respect to clustering, missing data and sample weighting, among other things. Social experimentation also has an elegant and abstract statistical theory very close to the survey research one. But it has a much shorter history of learning how to improve the implementation of social experiments; there has been less reflection on its own practices and less deliberate experimentation on its own practices. Nonetheless, great strides have been made over the last 30 years in identifying pitfalls to the quality implementation of social experiments. Without this self-critical work, McCall and Green would not have been able to make the points they did. And some progress has been achieved in learning how to circumvent many of the pitfalls, though as McCall and Green point out much still remains to be done. Thirty years from now we will probably have a theory of implementing experiments that is hodge-podge in form yet as useful in its results as the current theory of implementing sample surveys. Then, many of the objections to experiments will disappear.

However, McCall and Green are right in many other things. Experiments are a valuable resource only worth mounting when a causal issue is of the highest priority, when the relevant program theory is internally coherent and concordant with existing evidence, and when quality treatment implementation is likely. Also, applied developmental science will not progress through experiments alone, and certainly not through black box ones. Other kinds of issues count. They are also right that, even for the avowed task of testing causal relationships, experiments are no "gold standard". If one interprets a "gold standard" as a guarantee of uncontested inference, random assignment alone cannot achieve this. Interpreting its results depends on many other things—an unbiased assignment process, adequate statistical power, a consent process that does not distort the populations to which results can be generalized, and the absence of treatment-correlated attrition, resentful demoralization, treatment seepage and other unintended products of comparing treatments. Dealing with these matters requires observation, analysis and argumentation that are as likely to be needed 20 years from now as today, though the observations then are likely to be more targeted, the analyses more convincing and the arguments fewer, simpler and more empirically grounded. As I have tried to represent it here, the case for experiments is not that they are perfect. It is that they are better than their alternatives and that they will get even better as experience with them accumulates. But they are only better at one task. Yet identifying "what can (or does) work better than something else" is a centrally important task in both science and public policy.

References

Cook, T. D., Leviton, L. C., & Shadish, W. R., Jr. (1985). Program evaluation. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd edition) (pp. 699-777). New York: Knopf.

Cook, T.D. (2000). Towards a practical theory of external validity. In L. Bickman (Ed.) *Contributions to Research Design: Donald Campbell's Legacy*. Volume I. Newbury Park, CA: Sage.

Glazerman,S.,Levy ,D.M., & Meyers, D. (2003). Nonexperimental versus experimental estimates of earning impacts. *Annals of the American Academy of Political and Social Science*, *589*, 63-93.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181-1209.

Shadish, W.J., Cook, T.D. & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Continued from page 4

the decision process to which they were randomly assigned. If mediation produced better outcomes in a randomized trial, should it be mandated for everyone? Such a decision would clearly be an inappropriate generalization. If there is no difference, should we conclude that mediation is not beneficial for anyone? Surely not, because it might be quite beneficial for those who are interested in receiving that intervention. Participant choice may be crucial to the success of *either approach*.

The point here is that the effectiveness of the treatment may be quite different for those who actively seek out that treatment compared to those who are randomly assigned to it, and their reasons for participating in either research or service—need to be better understood to make good decisions about how the results can be generalized (Cook, 2000) and about whether the estimated effect sizes are a good reflection of "true" service effectiveness (statistical conclusion validity; Lipsey, 2000). Because of this, we suggest that selection processes and motivational factors leading to positive outcomes should be studied directly in social interventions, in addition to being "controlled" through randomization (Koroloff & Friesen, 1997). For the above example of divorce mediation, consider a design in which half of a sample of divorcing families are randomly assigned to either a) court determination or b) mediation, as in a traditional randomized trial, while the other half choose whether they want c) court determination or d) mediation. Such a design permits a direct examination of random assignment versus self-selected participation. Also, if variables thought to be related to the selection decision are measured in all groups, additional information can be obtained on the characteristics of families who are best served by the two treatments. Then such variables can be used to model selection bias and estimate the treatment effect over and above selection factors (Rossi, 1994) as well as guide advice to divorcing individuals regarding which procedure is likely to be most comfortable and effective for them.

Utilization of artificial selection criteria. A second problem is that traditional experimental research often uses a uniform and homogeneous participant population designed to minimize

extraneous participant factors and thereby maximize treatment-control group differences (Bickman, 1992). For example, the national evaluation of the Early Head Start (EHS) program originally requested that research sites limit enrollment to those families who had not used any other child and family services, including Head Start, within the past five years. This was not feasible in many urban areas in which very few eligible (low-income) families could be identified who had not used some kind of community service. More importantly, it was extremely unlikely that EHS services would be restricted to such families once the program was widely implemented, thus reducing the generality of these results to typical EHS service programs. Moreover, the policy decision to which the research is addressed (i.e., the "external validity criterion") is whether EHS is more effective than existing community services, not whether it is more effective than no services at all (which circumstance is no longer common). The more one attempts to eliminate participant factors in this way, the more limited and unique the sample and the less generalizable the results, especially to the real-world context in which a program will operate after the research demonstration.

Members of the same extended family were assigned to different groups…

Inappropriate use of individual-level random assignment. A third problem in the use of randomized designs is the reliance, sometimes inappropriately, on individual-level randomization. Although random assignment of other units, such as classrooms, schools, even cities has increased (e.g., Boruch & Foley, 2000), individual-level random assignment remains the most frequent in psychology. Two of the most common problems that result from such inappropriate random assignment of individuals are diffusion of treatment and resentful demoralization (Cook & Campbell, 1979).

Diffusion of treatment is the receipt of treatment by those in the no-treatment comparison group (Cook & Campbell 1979; Aiken & West, 1990). While relatively common (Orwin, Cordray, & Huebner, 1994), it has generally been ignored by many researchers and policy makers. Orwin et al. reviewed 14 alcohol treatment projects and found evidence that diffusion of treatment played a significant role in reducing the differences between treatment and comparison groups in 8 out of 14 studies (the most common of all the possible influences reviewed).

Resentful demoralization arises when members of the control group discover they did not receive the "special treatment" and either become disappointed ("demoralized") and have less positive outcomes than might be expected or overcompensate and do better (Aiken & West, 1990; Fetterman, 1982). Knowing that some participants are not receiving special services may also affect the service providers and lead them to change the way "services as usual" are delivered (Lam, Hartwell, & Jekel, 1994). Resentful demoralization, therefore, can act to either artificially increase *or* decrease treatment effect estimates in randomized designs; unfortunately it is difficult to predict *a priori* the direction or the extent of such bias.

Diffusion of treatment and resentful demoralization are separate problems, but they both are more likely to occur when individual participants are assigned to treatment vs. comparison groups in a context in which members of the groups will be in contact with one another. For example, in the Comprehensive Child Development Project (CCDP), ethnographers documented sharing of knowledge between individuals in the treatment group and those in the comparison group, a process participants called "passing it on" (Gilliam, Ripple, Zigler, & Leiter, 2000; McAllister, 1993). In the CCDP project, randomization was mandated at the individual level, but often within tightly knit neighborhoods and housing projects. In some instances, members of the same extended family were assigned to different groups, sometimes even when those family members shared the same residence. Moreover, the CCDP program focused on mobilizing an entire community to support its residents (Kagan & Weissbourd, 1994; H. Weiss & Halpern, 1991). It was impossible, for example, to prohibit control families from using a community drop-in center or to benefit from a community spirit of hope and regeneration that the treatment program attempted to foster.

Ideally, in situations such as this, communities, rather than individuals, might be randomly assigned to treatment vs. comparison groups (McCall, Ryan, & Plemons, 2003). Such designs have the added benefit of taking into account the fact that community and environmental factors may play important direct and indirect roles in contributing to program effectiveness (Boruch & Foley, 2000). Random assignment of larger units, such as communities, schools, or classrooms, is much more common in the fields of public health and education than in psychology; however, successful examples in a diverse array of fields can be found (Boruch & Foley, 2000). Admittedly, such "cluster random assignment" often requires more resources and larger numbers of individuals, has its own methodological disadvantages (e.g., comparability of communities), and may need different statistical techniques than those that psychologists most often use (Simpson, Klar, & Donner, 1995). But the advent of mixed model analysis and hierarchical linear modeling, for example, greatly increases the precision of our estimates of general treatment effects as well as the effects on individuals within the larger units for these types of designs (Boruch & Foley, 2000; Bryk & Raudenbush, 1992). Further, recent work suggests that careful matching and other procedures can greatly reduce the number of units that are required for statistically powerful designs (Bloom, 2003).

Further, when it is impractical to randomly assign these larger units, numerous quasi-experimental designs (e.g., when communities are selected, rather than randomly assigned), especially the non-equivalent control group design, can be employed. Such approaches, if well-designed and implemented, can yield effect size estimates comparable to those found in randomized studies (Heinsman & Shadish, 1996; Lipsey & Wilson, 1993). Unfortunately, they are often dismissed by psychological researchers or prohibited

by funders or policy makers because they are perceived as "inferior" for addressing questions of causality.

"Once randomized, always analyzed." Sometimes random assignment is preserved in data analyses at the exclusion of other analyses that would provide information that might be highly relevant to the ecological validity of the total findings. For example, national evaluators of major government-sponsored intervention demonstration projects typically include all participants who were *assigned* to the treatment and comparison groups in their analyses (an "intent to treat" model), including participants who refused or dropped out and actually did not receive any or all of the treatment. This strategy preserves random assignment and avoids the confound with selective dropout and the fear that the research and policy communities might reject the results of the entire evaluation because randomization was "compromised."

> It seems absurd to include people in the "treatment" group who never received program services...

On the other hand, from a conceptual perspective it seems absurd to include people in the "treatment" group who never received program services and to preclude analyses that would estimate how well the program worked specifically for participants who completed, or at least participated in the program to some minimum extent. Program participation in these types of services is rarely mandated; in the "real world" some people will enter and engage fully in services while others will not. Rather than asking "does the program work for potentially eligible applicants," an appropriate policy question is "does the program work for people who are willing and able to participate?" Participant retention is often a program goal, so retention rate may be considered an outcome variable. From an applied policy perspective, it is important to know the program's potential effectiveness and the types of participants and circumstances for which the program works best or not at all. Statistical techniques increasingly provide other strategies, such as instrumental variables estimation (Davidson & MacKinnon, 1993; Foster & McLanahan, 1996; Yoshikawa, Rosman, & Hsueh, 2001) and propensity scoring (Foster, 2003; Imbens, 2000; Rosenbaum & Rubin, 1983, 1985) that can model selection factors relating to program participation and unmeasured confounding. To adequately use such techniques, however, it is critical that researchers include baseline measurement of variables they believe may be related to selection. These variables are likely to go beyond what is typically available to persons doing selection modeling (e.g., demographic variables).

When substantial, selective, and differential dropout across treatment and comparison groups exists, bias is very likely present and there is no totally adequate correction. Each of the above approaches has it merits and limitations, so as many of them as are feasible and appropriate should be used to attempt to achieve a convergent conclusion.

Interpretation of results from randomized studies. From a traditional perspective, randomized trials are the only way to determine causality and to obtain an accurate estimate of effect size. But, as discussed above, a randomized trial may not produce unequivocal evidence for program effectiveness or accurate estimates of effect sizes. This problem relates generally to the issue of statistical conclusion validity; that is, the validity with which one concludes, based on available statistical tests, that an intervention did or did not have an effect (Lipsey, 2000). For example, it is possible, as in the divorce mediation example above, for a randomized comparison of mediation vs. court-determined custody to find no differences, yet one or both treatments might be effective over the other if parents choose (i.e., self-select) the treatment. Further, diffusion of treatment, resentful demoralization, and the inclusion of participants who did not receive any or all of the treatment might so dilute the treatment effect that the program is wrongly declared totally ineffective or as having only modest benefits.

> In some cases, the randomized trial produces a minimum estimate of effect size…

Regrettably, the results of randomized trials of behavioral interventions are often interpreted to represent the "true" effect size, as it is in many medical studies. But in the latter studies, especially drug trials, the potential influence the participant may have on the effectiveness of the treatment is minimal. In drug trials, for example, once people agree to participate, the effectiveness of a drug vs. a placebo is not differentially influenced much by the participants' motivation or perception of the specific treatment, in part because many drug trials use a double-blind design, which is typically impossible in many behavioral interventions.

Although the strength of the randomized trial is that it minimizes the confounding effect of participant characteristics and self-selection, it is shortsighted to use it as the "only" or "true" source of evidence in cases in which the effectiveness of the intervention may well be influenced substantially by the same participant characteristics and self-selection that the randomized trial is designed to minimize. In some cases, the randomized trial produces a minimum estimate of effect size, excluding all participant factors. That is useful information, but in situations in which participant characteristics are likely to play a role, it is also useful to know the maximum effect size when participant factors are included, and to recall in the interpretation that the "confounding" selection factors may be just as causal and necessary as the treatment program itself. And since publicly-funded services are never randomly assigned, a policymaker may be most interested in the effects of the program on those who choose the program, choose to stay in the program, or have certain characteristics,

especially those that can be used as eligibility criteria to target the program at specific types of participants.

No comparison group at all. As funders and policy makers continue to insist on "accountability" and "evidence-based programming," more programs created and operated by community agencies will need evaluations of their effectiveness. Not only do most such programs not have a randomly assigned control group, they do not have a comparison group at all. While many traditionalists would declare that there is no way to determine program effectiveness, newer approaches permit some estimation of outcomes using constructed comparisons. For example, one strategy is especially suited to developmental contexts in which an outcome variable is known to change with age, and children (or senior citizens) enter the treatment program at different ages and remain in treatment for different lengths of time (Bryk & Weisberg, 1977; McCall, Ryan, & Green, 1999). A simple pre-posttest design confounds developmental maturation with treatment effects. However, pretest scores (and/or scores from siblings) can be used to calculate an age function, which can predict individually for each participant using pretest score and time in program for that child's expected no-treatment comparison score at his or her age of actual posttest. The difference between actual posttest and predicted no-treatment scores represents a rather sensitive estimation of pre-post treatment effects controlling for no-treatment age changes, age and time in program, and individual differences without any independent comparison group.

Despite being in an era in which policy makers want "evidence-based programming," we frequently are embarrassingly left not knowing the "program" for which we have "evidence."

Experimenter-Controlled Uniform Treatment Manipulation

The third major feature that typifies gold standard experimental approaches is an experimenter-controlled treatment that is applied uniformly to all participants. This approach reduces error variance associated with treatment variability and specifies precisely the nature of the "cause" of any outcome effects. The assumption of uniformity of treatment is appropriate when the treatment of interest is singular and stable, such as the administration of a particular drug or medical procedure to treat a specific illness or disorder. However, this is rarely the case in behavioral and social research (Conrad & Conrad, 1994). In fact, research on social interventions almost never includes the years of preliminary work focused on understanding the treatment itself that commonly preface randomized clinical trials in medicine. Below we describe several problems that can result from applying methodologies that make this assumption in applied social research.

Lack of specification of treatment variables. A major putative advantage of the experimental method is that it allows researchers to know exactly what variables and in what amounts (i.e., the specific treatment or program) produced the outcomes that are observed. But this benefit may exist more in theory than in actual practice, especially in applied social research. For example, Brekke (1988, p. 946) writes:

> "Studies have shown that community support programs are more effective in treating the chronic mentally ill than traditional forms of aftercare. Yet an analysis of 33 controlled studies of community support programs reveals that almost no systematic empirical knowledge exists about their implementation, including the kinds of treatment they deliver, how they can be replicated, or what ingredients account for their success."

This is hardly an uncommon situation (Conrad & Conrad, 1994). Not much is known about the details of many major early childhood interventions, including some extremely prominent and influential interventions (e.g., the Perry Preschool Project, Healthy Families America, Even Start). Therefore, we frequently have little description, let alone empirical documentation, of the treatment or program that produced the observed benefits. Studies of program "dosage" which can offer at least some guidance to policy makers about the amount and intensity of services that are needed to produce an effect, are few and far between in the early childhood arena (Shonkoff & Phillips, 2000). Studies that have attempted to examine the impacts of varying levels of participation in programs have been harshly criticized because of problems related to selection bias. Recently, however, more approaches using statistical techniques such as propensity scoring have begun to explore dosage issues with greater sophistication (Foster, 2003; Hill, Brooks-Gunn, & Waldfogel, 2003). Such studies are sorely needed, because despite being in an era in which policy makers want "evidence-based programming," we frequently are embarrassingly left not knowing the "program" for which we have "evidence."

Changing programs and cross-site program variation. The urgency to conduct an outcome evaluation often means that the first (and sometimes only) cohort of participants and service providers are studied. But the nature of the treatment can change as both groups struggle with how to implement the treatment and informally learn what seems to work best. Campbell (1984) long advocated to "evaluate only proud programs" (p. 37), citing premature evaluation as one of the ten biggest mistakes of early applied social science. Ten years later, Cook continued to argue that social scientists are bypassing the necessary research aimed at understanding the relation between and among treatment and outcomes necessary to select a specific treatment before moving on to randomized outcome studies, which puts neophyte programs to a premature and unfair test (Cook, 1993; 1998).

One positive example of a large-scale outcome study that built in mechanisms to allow serious research on site-specific program functioning is the Early Head Start Demonstration Project. This evaluation, while retaining a congressionally-mandated, multi-sited evaluation using a randomized clinical trial design, also provided separate funding for each of the 17 participating sites to conduct local research, specifically focused on understanding the

mechanisms of change for that local program. Researchers in each site were funded to study variations in program implementation, individual differences in participants, and other contextual variables in an effort to determine what works, how, for whom, and under which circumstances. The EHS model comes close to one advocated by Campbell (1984), who called for the abolition of the large-scale multi-sited national evaluation in favor of funding that would allow "cross-validation" between multiple local versions of similar programs using different research designs; EHS formalized this approach into an integrated consortium.

Lack of specification of the comparison group. One also needs to specify the nature of services received by persons in the comparison, as well as the treatment, group. Frequently in human services research, "control" groups are defined, not as "no service" groups, but as "treatment as usual" groups, which may be called "comparison" groups to emphasize this circumstance (Gilliam et al., 2000). However, there is often little if any information about the "treatment as usual" services received by the comparison group, which a) increases the chance that there will be no significant difference in outcomes for treatment vs. comparison groups (e.g., if comparison group members received reasonably effective services that don't differ markedly from the new "treatment") and b) interferes with researchers' ability to understand the results (e.g., the treatment was better –or no different - than *what*?). Indeed, it may be difficult, depending on the service offered, to obtain a comparison group that receives substantially less service than the treatment group, thus obviating a meaningful and powerful comparison (Heinsman & Shadish, 1996). This was the case in the Comprehensive Child Development Program (Gilliam et al., 2000; Goodson, Layzer, St. Pierre, Bernstein, & Lopez, 2000; McCall et al., 2003; St. Pierre, Layzer, Goodson, & Bernstein, 1997), which failed to demonstrate many treatment-comparison group differences, perhaps because of high service rates in the comparison group (McCall et al., 2003).

Standardization of treatment vs. individualization of treatment. Perhaps the biggest challenge to methods designed for a single uniform treatment is the increasingly common program model that calls for individualizing services to fit family, child, or parent needs and characteristics. For example, a cardinal principle of good teaching, services, or therapy is that they be matched to the profile of strengths and limitations of individual participants. But this treatment strategy is in direct contrast to research methods that make vigorous attempts to insure uniformity by standardizing and monitoring the treatment (Sechrest, West, Phillips, Redner, & Yeaton, 1979). Schorr (common-purpose.org, 2001) suggests several ways that such "gold standards" in program interventions are poorly matched with traditional "gold standards" in research and evaluation, especially in relation to issues of treatment flexibility and individualization (Table 1, page 11).

Despite these obvious mismatches, major evaluation initiatives continue to act as if "the treatment" is, or should be, uniformly and consistently delivered to all participants. For example, one of the principal tenets of CCDP programs was that families would identify their own needs, goals, and services, so *by design* different families received different types and dosages of services based on individual needs. Such "family support programs," by definition, are not one-size-fits-all treatment approaches, yet the national study (St. Pierre et al., 1997) evaluated the CCDP program as if every family received the same treatment. That is, the completion of schooling outcome, for example, was assessed on all families in the treatment group even though only a small percentage of those families identified this as a need and made any attempt to receive training or education. Given the many possible needs, goals, and services in this "comprehensive" program, only small portions of the treatment sample received services designed to produce any single outcome. So it should come as no surprise that CCDP was reported to have achieved very few of its intended outcomes when analyzed in the total sample without regard to who selected a particular goal and relevant service and who did not (Gilliam et al., 2000; McCall et al., 2003).[1]

These problems suggest that research that conscientiously reflects the complexities of programs as they are delivered in the field is of utmost importance to understanding and improving service programs. Guralnick (1997) called this "second generation research," and while its importance has been increasingly recognized, there remains relatively little social intervention research that grapples directly with these issues.

CONCLUSIONS

If a major goal of science is to contribute to human welfare (McCall, 1996), then scientific research on social interventions should address both cause-effect questions as well as questions of ecological-validity and practical importance (McCall & Groark, 2000). Respect for both approaches and tasks should rest on the quality of the work within each context, not on a categorical preference for one type of research methodology over the other.

However, despite calls by numerous methodologists (e.g., Campbell, 1984; Cook, 1998; Orwin et al., 1994; and many others) for more judicious use of research methods, policy makers and many researchers continue to prefer certain traditional approaches, sometimes exclusively and unquestioningly. Researchers need to convey a more balanced message about the relative merits and limitations of different approaches. Specifically, researchers, practitioners, and policy makers who are involved in the study of social interventions should:

1. Adopt program development procedures that place greater emphasis on specifying the theoretical and empirical nature of expected relations between treatment approach, treatment activities, and outcomes (e.g., use "theory-based evaluation," "theories of change," or "logic model" approaches; Chen & Rossi,

[1]In fairness to the evaluator (Apt Associates), the funder issued separate contracts for the outcome evaluation and the national management information system, which contained the goals and service utilization records of each family. No provisions were made to integrate the two databases, so it was not possible to relate individual goals and service utilization to specific outcomes.

Continued on page 12

Table 1.

Mismatch Between the "Gold Standard" of Interventions and the "Gold Standard" of Evaluations

Attributes of Effective Interventions	Attributes Associated with Traditional "Evaluability"
Significant front-line flexibility within established quality parameters	Intervention standardized; discretion minimized
Evolving – in response to experience and changing conditions	Intervention constant over time
Intervention/program design reflecting local strengths, needs, preferences	Intervention centrally designed, uniform across sites
Intake/recruitment into program under local control within broad parameters	Intake/recruitment centrally designed to permit random assignment
Multiple components respond to children in family, peer, & neighborhood context	Single factor, single sector
Interactive components take into account health, social, educational needs	Components clearly separable
Emphasize continuing respectful relationships, other hard-to-measure attributes	Readily measured inputs
Implementers "believe in" the intervention	Value-free implementation

From Schorr (common-purpose.org, 2001) with permission.

1983; C. Weiss, 1995), recognizing that society needs some questions addressed even without a theoretical basis.

2. Understand that methods created primarily to establish internal validity in laboratory contexts typically cannot be easily or directly utilized in the context of applied research and evaluation without modification and supplementation with other methods (guidelines for conducting such research can be found in Groark & McCall, in press; McCall, Green, Strauss, & Groark, 1998).

Quarrels over which method represents 'the gold standard' make no more sense than arguing about whether hammers are superior to saws.

3. Use and value fairly and accurately the contributions and limitations of a variety of methodological approaches (e.g., "planned critical multiplism;" Shadish, 1986). Be creative and employ one or more designs that maximize both internal and external validity, including those that allow some self-selection into treatment, that introduce flexibility to match treatment with participants, and that employ randomization at levels other than the individual when appropriate.

4. Employ designs that go beyond simple reliance on treatment vs. control group comparisons, such as within-treatment analyses of varying levels of program involvement or service delivery and designs that focus on developing and testing theories of change in systematic ways (C. Weiss, 1997).

5. Place greater emphasis on studying program implementation before studying outcomes to learn what aspects of the original treatment program can be realistically implemented and seem to be effective for which kinds of participants. This initial process phase would allow refinements in the treatment and help to determine which treatment(s) can and should be evaluated for outcome effectiveness.

6. Measure and study in greater detail the treatment(s) as actually implemented and the services actually received by participants in both the treatment and comparison conditions, preferably with both qualitative and quantitative assessments of treatment administration.

7. Recognize and allow for individualized treatments when appropriate, describe and study the individualization process, model treatment selection parameters to use as control factors, and relate differences in participant treatment usage to differences in specific outcomes (Friesen & Koroloff, 1990; Green, Rodgers, & Johnson, 1999; Hill et al, 2003).

8. Utilize data analysis procedures that explore selection factors, individual growth trajectories, and complex relations among treatment variables and between treatment and outcome variables outside of randomized designs, including growth curve modeling, propensity scoring, mixed-model analysis, and instrumental variable techniques (e.g., Foster & McLanahan, 1996; Rosenbaum & Rubin, 1985; Raudenbush & Liu, 2001). To maximize the usefulness of these approaches, researchers will need to carefully consider their assumptions and measure *a priori* those variables that may be related to selection and participation in treatment.

This general approach will take longer and cost more (Bickman, 1989), require a closer partnership between program and evaluation personnel (Green & McAllister, 1998) and thus potentially sacrifice the often-desired "independent" or "outside" evaluation (Gaventa, Creed, & Morrissey, 1998), and increase the complexity of multi-sited projects in which "the program" is quite different at different sites. Moreover, it should be noted that programs that engage in a research process that explicates and measures detailed program implementation are likely to be fundamentally different from programs that do not do so, thus introducing an external validity concern. However, in our experience, a single, uniformly-administered treatment within one or across several sites is largely a myth; these procedures recognize this reality and try to describe and study it. Such research can help to guide better implementation of programs that do not have the resources or expertise to engage in ongoing intensive evaluation themselves.

The potential of our discipline to contribute to human welfare, especially of children, youth, and families, is great. But this potential will not be achieved if we are constrained to only using one vs. another methodological approach. Research methods are tools that must match the scientific, practice, and policy tasks, and the research question and intervention should dictate the method, not the reverse. We are more likely to maximize our contribution if we broaden our methodological value system to recognize the benefits and limitations of all methods. Schorr and Yankelovich (2000) recently concurred: "Many new approaches now are becoming available for evaluating whether complex programs work…. Quarrels over which method represents 'the gold standard' make no more sense than arguing about whether hammers are superior to saws. The choice depends on whether you want to drive in a nail or cut a board" (p. B7).

Notes

This paper is supported in part by an Urban Community Services Program Grant No. P252A50226 from the U.S. Department of Education to Robert B. McCall, Ph.D., and by a grant to McCall and Carey Ryan from the Howard Heinz Endowment. McCall's address: Office of Child Development, University of Pittsburgh, 400 N. Lexington Avenue, Pittsburgh, PA 15208 (mccall2@pitt.edu).

Why We Need More, Not Fewer, Gold Standard Evaluations

Phoebe Cottingham, *Commissioner of Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education*

McCall and Green correctly note that random assignment is not the only valuable method of program evaluation and that experimental methods can be applied inappropriately. I agree with these two points. However, I disagree on two points. First, their claim that there is too much emphasis on randomized controlled trials (RCTs) among developmental psychologists who study real-world social issues is questionable. Second, their championing of newer non-experimental techniques ignores compelling evidence that these methods are not as good as experimental designs at determining program impacts.

First, are RCTs so dominant? In 2003, 112 empirical articles were published in the SRCD journal *Child Development*. Of those, 30 clearly addressed "real-world social issues," such as effects of welfare reform on children or the effectiveness of parenting workshops for single, low-income mothers. Of those 30, only 3 were program evaluations that addressed the impact of social policies on individual outcomes using RCTs. The overwhelming majority of studies of real-world issues used tools among McCall and Green's list of alternatives to random assignment, such as correlational methods, hierarchical linear modeling, and non-equivalent control groups. Hence, in the field of child development research, the problem appears to be the underutilization of RCTs, not the overemphasis decried by McCall and Green, admitting that most of this work is not program evaluations.

More to the central contention in the article is the emphasis on how experimental studies may produce "misleading" findings. McCall and Green define RCTs as "experimenter-controlled manipulations uniformly applied to all participants under rigorously controlled conditions." This is an idealized conception of RCT's. Field experiments on real-life problems rarely if ever are designed with the expectation of a uniform intervention applied under controlled conditions. Indeed, variation in program implementation and participant involvement, attrition, diffusion of treatment and other such issues are taken for granted and analyzed. They are the bread and butter of sophisticated evaluators who work with RCTs. McCall and Green argue that these issues are problems of RCTs. However, these phenomena exist whether the evaluation design is experimental or not.

The advantage of experimental designs is they produce a true counterfactual in the randomized control group. One can isolate changes due to the availability of the treatment – the main question — from these other sources of variation in behavior. Without the randomization process, one has to guess what might be the selection factors that influence behavioral responses to availability of treatment. Randomization puts everyone in the same place and assures that the same distribution of possible selection factors exist in both the experimental and the control group. One may never know what all these factors are, but one has confidence that they are distributed by chance across the two groups. This allows for the "control" of the selection factors. Once the assignment is made, these selection factors will continue to influence behavior, but we can have confidence that differences detected between the two groups formed by random assignment are due to the availability of treatment.

Furthermore, McCall and Green conclude that the experimental method is just another tool in the toolbox for determining program effects. In doing so, they ignore a growing body of evidence that points in a different direction. A recent report from an ongoing review of design replication studies concluded that no single quasi-experimental method reliably produces answers that are close enough to the answers obtained from an experimental design (Steven Glazerman, Dan M. Levy, and David Myers, "Nonexperimental versus Experimental Estimates of Earnings Impacts", *The Annals of the American Academy of Political and Social Science*, September 2003). The review looked at 12 independent studies by researchers who carefully tested the new non-experimental methods. Each study used actual field experiment data as the study base and employed statistical methods such as propensity scoring or other matching methods to simulate what would be obtained as impact estimates if non-experimental methods had been used instead of the experimental method. Sometimes the researcher drew a comparison group from a national dataset or drew from control groups for other sites. Whether the non-experimental method was regression, matching, or difference in difference, these studies found a lack of consistent congruence between the non-experimental estimates and the experimental estimates. In one case involving studies of welfare reform, results from comparison groups formed using sophisticated statistical procedures had a 40% chance of being far off from the results obtained with random assignment. (Howard Bloom, Charles Michalopoulos, Carolyn Hill, and Ying Lei, *Can non-experimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?* MDRC, New York, June 2002).

Researchers and evaluators interested in determining the effectiveness of real-life social and education programs must understand the limitations of the non-experimental methods. The nation cannot afford to construct social policy around methods that can generate wrong answers. We need to continue to explore alternatives to randomization with the hope of identifying the conditions under which non-experimental methods generate acceptable approximations of results that can be obtained from experiments. However, for now, the best and most reliable method available for determining what works is the experiment, which therefore can be considered a gold standard.

Don't Throw out the Baby with the Bathwater: Incorporating Behavioral Research into Evaluations

Jeanne Brooks-Gunn, *Columbia University*

While McCall and Green are correct in asserting that Randomized Control Trials (RCT's) have, like all other methods, some limitations, they are more likely to address the problem of unobserved differences between groups (or selection bias) than other methods (see Heckman & Smith, 1997 for a discussion of limitations of RCTs). I wish to make two points in this essay, in the spirit of reconciliation among those who favor experimental and those who prefer non-experimental methods. First, RCTs themselves could be used more creatively to provide more information on topics such as differential effects of implementation, dose of treatment, curriculum approaches, service delivery, diffusion, and responsivity to treatment. Second, the judicial use of non-experimental methods can provide 'upper' and 'lower' bounds of effect sizes and are critical to estimate effects of policies and behaviors that are not amenable to randomization (divorce, maternal employment, teenage parenthood).

First, more nuanced information about treatment effects as well as theory can be derived from RCTs. Several examples are drawn from the early childhood education (ECE) literature, which has the largest number of RCTs of the child and adolescent programs to enhance well-being (Brooks-Gunn, 2003). The most obvious elaboration is to test not one treatment against a control group (in social science the control group might be more accurately termed a follow-up group, since members of this group receive whatever treatment is currently available in a community or school). Ramey and colleagues tested whether home visiting alone or home visiting in conjunction with center-based care during the preschool years made a difference. They found that their home visiting intervention was not effective unless paired with the center program (Wasik et al, 1990; see also Schweinhart et al, 1986, for a comparison of three different preschool curricula). The Early Head Start National Evaluation compared three types of programs in their 17 site RCT—those offering home visiting, center care, or both (although children were not assigned via randomization to the 3 types of programs, effect sizes across the three program types were compared; this hybrid-design resulted in larger effects for center-based and center- and home-based programs than for the solely home-based programs; Love, et. al., 2003).

Another elaboration involves dose-related analyses; again, these are not strictly experimental, unless children are assigned via randomization to receive different intensities of treatment. Propensity score matching techniques with many control variables render treatment and control groups quite similar (Rosenbaum & Rubin, 1985; in this ECE case, treatment groups are created via the amount of intervention actually received). Looking at effects at age 8 of a center- and home-based intervention in the first 3 years of life, Hill and her colleagues reported an 11- to 14-point IQ benefit for (heavier low birth weight) children who received at least 400 days of center-based intervention compared to the children in the control group, and about a 6- to 8-point advantage for children who received at least 350 days (Hill, et. al., 2003). Another approach used with the Comprehensive Child Development Programs involves comparing sites with higher or lower levels of services (akin to the hybrid – design used with Early Head Start; Love, Brooks-Gunn, et. al., 2004).

Controls for take-up rate of a treatment also are available now, in the form of Treatment on Treated (TOT) analyses, which take into account the fact that not all families who are randomized into a treatment group actually receive the treatment. These estimates are usually reported along with the Intent to Treat (ITT) analyses (see Love, et. al., 2003 for an example). If most families receive the treatment, then the estimates for ITT and TOT are quite similar, as in the case of many ECE programs (for an example of a program where take-up rates were expected to be much lower, and were, see the Moving to Opportunity evaluation; Leventhal & Brooks-Gunn, 2003).

Possible effects of diffusion also can be examined, either by comparing the effect sizes seen for ECE from the 1970's to the present (given that many fewer mothers of young children were working 30 years ago, so that most children in the control group were not receiving any center-based services). Or, using propensity score matching procedures, it is possible to match treatment group families to control group families based on the type of child care that the control group families are using. Comparisons, then, are made between control group children who received no outside-the-home care and those treatment group children who would not have received outside care if the intervention had not existed (Hill, et. al., 2002). Likewise, children in the control group who did receive center-based care in the community were compared to children in the treatment group who would have received such care in the absence of the intervention.

Second, some non-experimental approaches are likely to minimize selection bias. The most obvious involves what are often called 'naturally occurring experiments,' since they involve an exogenous shock. A recent example is the introduction of casinos. During a longitudinal study in the Great Smoky Mountains, in one area, members of a Native American-Indian tribe received cash and some members also became employed, neither of which were true of comparably poor non-Native Americans in the same region; well-being increased for the Native American children but not the other children who were initially poor (Costello, et. al., 2003). Other designs include the 'instrumental variable' and the 'difference-in-difference' approaches. The former seeks to identify a variable that influences the likelihood of receiving a treatment but not the outcome of interest while the latter identifies two groups and assesses them at least twice, once before the 'treated' group experiences the treatment and once post-treatment, with a control group being assessed at the same time points (but does not receive the treatment in the interim). Yet another procedure involves statistical matching, as described earlier (propensity score procedure). Other strategies capitalize on longitudinal designs where individual fixed effects may be modeled or where panel studies of families allow comparisons among siblings or between twins, to control in the first case for some shared environmental effects and in the second case for some genetic effects. My point is that these methods, which are much more common in sociology, demography, and economics, have much to offer in the study of children, youth and families (Thornton, et. al., 2001).

References

Brooks-Gunn, J. (2003). Do you believe in magic?: What we can expect from early childhood intervention programs. *Social Policy Report, Society for Research in Child Development, 17* (1). http://www.srcd.org/spr17-1.pdf.

Costello, E.J., Compton, S.N., Keeler, G., & Angold, A. (2003). Relationships between poverty and psychopathology: A natural experiment. *The Journal of the American Medical Association*, *290*, (15), 2023-2029.

Heckman, J.J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies, 64*, 487-535.

Hill, J., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participations in an early intervention for low-birth weight premature infants. *Developmental Psychology, 39* (4), 730-744.

Hill, J., Waldfogel, J. & Brooks-Gunn, J. (2002). Assessing the differential impacts of high-quality child care: A new approach for exploiting post-treatment variables. *Journal of policy Analyses and Management, 21* (4), 601-627.

Leventhal, T. & Brooks-Gunn, J. (2003). Moving to Opportunity: An experimental study of neighborhood effects on mental health. *American Journal of Public Health, 93* (3), 1576-1582.

Love, J.M., Constantine, J., Paulsell, D., R., Boller, K., Ross, C., Raikes, H., Brady-Smith, C., Brooks-Gunn, J. (2004) The role of Early Head Start programs in addressing the child care needs of low-income families with infants and toddlers: Influences on child care use and quality. Washington, DC: U.S. Department of Health and Human Services. Retrieved from http://www.mathmatica-mpr.com/PDFs

Love, J. M., Kisker, E. E., Ross, C. M., Schochet, P. Z., Brooks-Gunn, J., Paulsell, D., Boller, K., Constantine, J., Vogel, C., Fuligni, A. S., & Brady-Smith, C. (2002*). Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start*. Washington, DC: U.S. Department of Health and Human Services. Retrieved from http://www.mathematicampr.com/3rdLevel/EHSTOC.HTM.

Rosenbaum, P.R. & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33-38.

Schweinhart, L., Weikart, D., & Larner, M. (1986). Consequences of three preschool curriculum models through age 15. *Early Childhood Research Quarterly, 1,* 15-45.

Thornton, A., (2001). (Editor). *The well-being of children and families: Research and data needs.* Ann Arbor, MI: University of Michigan Press.

Wasik, B.H., Ramey, C.T., Bryant, D.M., & Sparling, J.J. (1990). A longitudinal study of two early intervention strategies: Project CARE. *Child Development, 61* (6), 1682-1695.

On Randomized Trials and Bathwater: A Response to Cottingham and Brooks-Gunn

Robert B. McCall, *University of Pittsburgh,* Beth Green, *NTC Research*

We believe both the commentaries by Cottingham and Brooks-Gunn (this issue) are useful in complementary ways.

First, Cottingham disputes our assertion that randomized control trials (RCT) are over emphasized in developmental research, indicating that more applied articles published in *Child Development* employ non-experimental than experimental methods. Cook (2000) makes a similar assertion for educational program evaluation.

We did not argue that non-experimental approaches were not *used*, but that they were *undervalued* relative to RCTs for their potential contributions.

Cottingham also argues that we ignore compelling evidence that non-experimental methods are not as good as experimental designs at determining program effectiveness. Undoubtedly, this depends on the particular methodologies and circumstances involved, and scholars vary in their conclusion on this issue (e.g., for different views from those of Cottingham, see Heinsman & Shadish, 1996; Lipsey, 2000; Lipsey & Wilson, 1993).

Regardless of the truth of the assertion, the claim itself illustrates one of our major points. We do not say non-experimental methods can be or should be a substitute for experimental methods. Experimental and non-experimental methods have different purposes, strengths, and limitations. RCT attempt to maximize internal validity, for example, whereas many non-experimental approaches attempt to maximize external validity. Both kinds of information are needed, and both approaches have their limitations.

Some new strategies (e.g., propensity scores) attempt to compensate for some of the limitations of non-experimental methods (e.g., participant selection bias). Conversely, we have suggested that RCT can produce biased and inaccurate estimates of treatment effects, usually to an unexamined and thus unknown extent, and some non-experimental approaches (e.g., participant selection versus random assignment) attempt to compensate for some of these limitations of RCT (e.g., participant motivation). Brooks-Gunn provides a highly useful list of strategies that can help minimize the limitations of both of these approaches. This is why we do not favor throwing the experimental baby out with the bathwater, but neither should the baby go unwashed and not benefit from the strengths of other research strategies, some of which are strong in precisely the ways that the randomized trial is potentially weak.

The potential bias and inaccuracy in RCT designs also make it unwise to hold up results from RCT designs as the standard (i.e., the "true" or maximum effect size) against which all other methods are compared. As Brooks-Gunn suggests, it is not necessarily clear which estimate of treatment effects is the most accurate, and we would add that accuracy is relative to the purpose and circumstances of the research (randomized trials or non-randomized conditions that are likely to prevail when the intervention is provided as a routine service.)

For example, Datta (2003) reports that the randomized control evaluation (Millsap, Chase, Obeidallah, Perez-Smith, Brigham, & Johnston, 2000) of the "Comer program" to improve schools (Comer, Haynes, Joyner, & Ben-Avie, 1996) showed no difference between Comer and control schools. However, schools in the treatment group varied in the extent to which they faithfully implemented the Comer program, presumably reflecting variations in motivation and enthusiasm for the program assigned to them. Analogously, some schools randomly assigned to the no-treatment control group did not obediently sit on their hands, and were also motivated to implement on their own principles of the Comer program (as a result of "resentful demoralization"?). The extent to which schools adopted the Comer program was found to be related to benefits for children in both groups.

Now, which is the "true effect size" – the no-difference, randomized result or the non-randomized, self-selected effect in both groups of fidelity to the Comer program? From a traditional scientific standpoint, is the self-selection result possibly due to the school's motivation to implement Comer principles? Probably. Even to the extent that the Comer program per se is not effective? We don't know. From the standpoint of an applied professional funder or policy maker, do we dismiss the self-selected results? No, because in practice only schools that are enthused about the program will likely choose the program, implement it well, and succeed at it. From this practice and policy perspective, the self-selection result is the more relevant ("true"?) effect size. Should more schools be funded and supported in implementing the Comer program? Yes. Applied research, practice, and policy are best served by using and valuing a variety of methods for the diverse purposes they serve, recognizing the strengths and limitations of each approach.

References

Comer, J. P., Haynes, N. M., Joyner, E. T., & Ben-Avie, M. (Eds.) (1996). *Rally the whole village: The Comer process for reforming education*. New York: Teachers College Press.

Cook, T. D. (2000). Toward a practical theory of external validity. In L. Bickman (Ed.), *Validity & social experimentation*, (pp. 3-44). Thousand Oaks, CA: Sage.

Datta, L.E. (2003). Avoiding unwarranted death by evaluation. *The Evaluation Exchange, 9*(2), 5.

Heinsman, D. T., & Shadish, W.R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154-169.

Lipsey, M. (2000). Statistical conclusion validity for intervention research: A significant ($p < .05$) problem. In L. Bickman (Ed.), *Validity & social experimentation*, (pp. 101-120). Thousand Oaks, CA: Sage.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, education, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Millsap, M. A., Chase, A., Obeidallah, D., Perez-Smith, A., Brigham, N., & Johnston, K. (2000). *Evaluation of Detroit's Comer Schools and Families Initiative*. Cambridge, MA: Abt Associates.

About the Authors

Robert B. McCall, Ph.D., is Co-Director of the University of Pittsburgh Office of Child Development and Professor of Psychology. He received his B.A at DePauw University and M.A. and Ph.D. at the University of Illinois. McCall has published on infant mental development, age changes in general mental performance, the prediction of later IQ, early childhood care and education, developmental research design and analysis, science communications through the media, and university-community partnerships. He has been an Associate Editor of *Child Development* and on the editorial boards of 10 other journals, and he was a Contributing Editor, monthly columnist, and feature writer for *Parents* magazine. He has received awards for contributions to research, public service, media, and policy from the American Psychological Association, the Society for Research in Child Development, the American Academy of Pediatrics, the National Council on Family Relations, and the University of Pittsburgh.

Beth Green is a Senior Research Associate at NPC Research, Inc. She received her Ph.D. in Social Psychology from Arizona State University in 1993, with an emphasis in applied research methods, and since then has been involved in conducting program- and policy-relevant research to support positive outcomes for low income and at-risk children and families. Her experience includes designing, implementing, and managing evaluations of a diverse array of programs including early childhood prevention and intervention, family support, drug and alcohol abuse prevention, community development, child welfare, and family treatment drug courts. Dr. Green is a member of the Early Head Start Research Consortium, a multi-site research collaborative conducting the national longitudinal evaluation of the Early Head Start program. In all of her work, she is dedicated to producing useful information through mixed method, theory-driven, collaborative approaches, and she has written and presented widely on these topics.

Jeanne Brooks-Gunn is the Virginia and Leonard Marx Professor at Columbia University's Teachers College and the College of Physicians and Surgeons. She directs the National Center for Children and Families at Teachers College and Columbia University's Institute of Child and Family Policy.

Thomas D. Cook is the Joan and Serepta Harrison Chair in Ethics and Justice at Northwestern University where he is also Professor of Sociology, Psychology, and Education and Social Policy and a Faculty Associate at the Institute for Policy Research. He is interested in evaluation, social experimentation and quasi-experimentation, and in contextual factors that influence adolescent development. He is a fellow of the American Academy of Arts and Sciences and a Trustee of the Russell Sage Foundation.

Phoebe H. Cottingham, Ph.D., was appointed Commissioner of Education Evaluation and Regional Assistance at the Institute of Education Sciences, U.S. Department of Education in August 2003. Previously she was a senior program officer in the Domestic Public Policy Program at the Smith Richardson Foundation, Inc. in Westport, Connecticut. Cottingham developed strategies to foster high-quality evaluations of interventions and policies in education, welfare, and family policy. She also helped develop new approaches to assessing intervention evidence and disseminating findings in social and education policy areas. Earlier, Cottingham was an associate director in the Equal Opportunity Program at the Rockefeller Foundation (from 1979 to 1996) after serving in research and teaching positions at Vanderbilt University, University of Pennsylvania, and the University of California at Berkeley and at Los Angeles. Cottingham obtained her Ph.D. in economics from U.C.-Berkeley where she specialized in urban economics and economic development.

References

Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. *Evaluation Review, 14*(4), 374-390.

Bickman, L. (1989). Barriers to the use of program theory, *Evaluation Practice*, *12*, 387-390.

Bickman, L. (1992). Designing outcome evaluations for children's mental health services: Improving internal validity. In L. Bickman (Ed.), *Evaluating mental health services for children* (pp. 57-68). San Francisco: Jossey-Bass.

Bloom, H. S. (2003). Sample design for an evaluatin of the Reading First program. U.S. Department of Education, Institute of Education Sciences.

Boruch, R. F. & Foley, E. (2000). The honestly experimenting society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), *Validity & social experimentation*, (pp. 3-44). Thousand Oaks, CA: Sage.

Brekke, J. S. (1988). What do we really know about community support programs? Strategies for better monitoring. *Hospital and Community Psychiatry*, *39*(9), 946-52.

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioral research: applications and data analysis methods*. Newbury Park, CA: Sage.

Bryk, A. S., & Weisberg, H. I. (1977). Value added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, *1*(2), 127-155.

Campbell Collaboration Library, http://www.campbellcollaboration.org/

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*, 409-429.

Campbell, D. T. (1984). Can we be scientific in applied social science? *Evaluation Studies Review Annual, 9*, 26-48.

Cauce, A. M., Ryan, K. D., & Grove, K. (1998). Children and adolescents of color: Where are you? Participation, selection, recruitment, and retention in developmental research. In V. C. McLoyd and L. Steinberg (Eds.), *Studying minority adolescents: Conceptual, methodological, and theoretical issues*. New Jersey: Lawrence Erlbaum.

Chen, H., & Rossi, P. (1987). The theory-driven approach to validity. *Evaluation and Program Planning, 10*, 95-103.

Chen, H., & Rossi, P. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7*(3), pp. 283-302.

Conner, R. F. (1977). Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation Quarterly 1(2)*, 195-244.

Conrad, K. J., & Conrad, K. M. (1994). Reassessing validity threats in experiments: Focus on construct validity. In K. J. Conrad (Ed.), *New Directions in Program Evaluation*, *63*, 5-26.

Cook, T. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest and A. G. Scott (Eds.), *New Directions in Program Evaluation, 57*, 23-78.

Cook, T. (1998). A discussion with Tom Cook. Roundtable presentation, American Evaluation Association Annual Meeting, November 1998, Chicago, Ill.

Cook, T. D. (2000). Toward a practical theory of external validity. In L. Bickman (Ed.), *Validity & social experimentation*, (pp. 3-44). Thousand Oaks, CA: Sage.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. Educational Evaluation and Policy Analysis, *24*, (3), 175-199.

Cook, T., & Campbell, D. T. (1979). *Quasi-experimentation*. Chicago: Rand-McNally.

Datta, L. E. (1994). Paradigm wars: A basis for peaceful coexistence and beyond. In C. S. Reichardt and S. F. Rallis (Eds.), The qualitative-quantitative debate: New perspectives (pp. 53-70). *New Directions for Program Evaluation Monograph* (No. 61). San Francisco, CA: Jossey-Bass.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Dunford, F. W. (1990). Random assignment: Practical considerations from field experiments. *Evaluation and Program Planning*, *13*, 125-132.

Fetterman, D. M. (1982). Ibsen's baths: Reactivity and insensitivity (A misapplication of the treatment-control design in a national evaluation). *Educational Evaluation and Policy Analysis*, *3***,** 261-279.

Foster, E. M. (2003). Is more treatment better than less? An application of propensity score analysis. *Medical Care*, *41*(10), 1183-1192.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods*, *1*, 249-260.

Friesen, B.J., & Koroloff, N. M. (1990). Family-centered services: Implications for mental health administration and research. *Journal of Mental Health Administration, 17,* 13-25.

Gaventa, J., Creed, V., & Morrissey, J. (1998, winter). Scaling up: Participatory monitoring and evaluation of a federal empowerment program. *New Directions for Evaluation*, *80*, 81-94

Gilliam, W. S., Ripple, C. H., Zigler, E. F. & Leiter, V. (2000). Evaluating child and family demonstration initiatives: Lessons from the Comprehensive Child Development Program. *Early Childhood Research Quarterly*, *15*, 41-59.

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

Green, B., & McAllister, C. (1998, March/April). Theory-based, participatory evaluation: A powerful tool for evaluating family support programs. *Zero-to-Three*, 30-36.

Green, B., Rodgers, A., & Johnson, S. (1999). Understanding patterns of service delivery and participation in community-based family support programs. *Children's Services: Social Policy, Research, and Practice*, *2*(1), 1-22.

Groark, C. G., & McCall, R. B. (in press). Integrating developmental scholarship into practice and policy. In M. H. Bornstein and M.E. Lamb (Eds.), *Developmental psychology: An advanced textbook, 5th edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Guralnick, M. J. (1997). Second-generation research in the field of early intervention. In M. J. Guralnick (Ed.), *The Effectiveness of Early Intervention* (pp. 3-22). Paul Brookes Publishing: Baltimore Md.

Haskins, R. (2001, June 23). What policy makers need from family researchers. Presented at the Third Annual Summer Institute, Family Research Consortium III, Public Policy, Socioeconomic Disadvantage, & Child Development. South Lake Tahoe, CA.

Hedrick, T. E. (1994, Spring). The quantitative-qualitative debate: Possibilities for integration. In C. S. Reichardt and S. F. Rallis (Eds.), The qualitative-quantitative debate: New perspectives (pp. 45-52). *New Directions for Program Evaluation Monograph* No. 61, San Francisco: Jossey-Bass.

Heinsman, D. T., & Shadish, W.R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154-169.

Hill, J. L. Brooks-Gunn, J. B., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*(4), 730-744.

House, T. E. (1994, Spring). Integrating the quantitative and qualitative. In C. S. Reichardt and S. F. Rallis (Eds.), The qualitative-quantitative debate: New perspectives (pp. 13-22). *New Directions for Program Evaluation Monograph* (No. 61) San Francisco: Jossey-Bass.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*, 706-710.

Kagan, L., & Weissbourd, B. (1994). Toward a new normative system of family support. In L. Kagan and B. Weissbourd (Eds.), *Putting families first: American's family support movement and the challenge of change*. San Francisco: Jossey-Bass.

Koroloff, N.M., & Friesen, B.J. (1997). Challenges in conducting family-centered mental health services and research. *Journal of Emotional and Behavioral Disorders*, *5*(3), 130-137.

Lam, J. A., Hartwell, S. W., & Jekel, J. F. (1994). "I prayed real hard, so I know I'll get in": Living with randomization. In K. J. Conrad (Ed), *New Directions in Program Evaluation*, *63*, 55-66.

Lipsey, M. (2000). Statistical conclusion validity for intervention research: A significant ($p < .05$) problem. In L. Bickman (Ed.), *Validity & social experimentation*, (pp. 101-120). Thousand Oaks, CA: Sage.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, education, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

McAllister, C. (1993). The impact of the CCDP on communities in CCDP service areas: Family Foundation's Comprehensive Child Development Program, Ethnographer's Report #10. Washington, D.C: Administration for Children, Youth, and Families, Department of Health and Human Services.

McCall, R. B. (1996). The concept and practice of education, research, and public service in university psychology departments. *American Psychologist, 51*, 379-388.

McCall, R. B., Green, B. L., Strauss, M. S., & Groark, C. J. (1998). Issues in community-based research and program evaluation. In I. E. Sigel and K. A. Renninger (Eds.), *Handbook of Child Psychology, Vol. 4* (*5th Edition*) (pp. 955-997). New York: Wiley.

McCall, R. B., & Groark, C. J. (2000). The future of child development research and public policy. *Child Development*, *71*, 187-204.

McCall, R. B., Groark, C. J., & Nelkin, R. P. (in press). Integrating developmental scholarship and society: From dissemination and accountability to evidence-based programming and policies. *Merrill-Palmer Quarterly*.

McCall, R. B., Ryan, C. S., & Green, B. L. (1999). Some non-randomized constructed comparison groups for evaluating age-related outcomes of intervention programs. *American Journal of Evaluation*, *2* (20), 213-226.

McCall, R. B., Ryan, C. S., & Plemons, B. W. (2003). Some lessons learned on evaluating community-based, two-generation service programs: The case of the Comprehensive Child Development Program (CCDP). *Journal of Applied Developmental Psychology*, *24*, 125-141.

McCurdy, K. & Daro, D. (2001). Parent involvement in family support: An integrated theory. *Family Relations, 50,* 113-121.

Orwin, R. G., Cordray, D. S., & Huebner, R. B. (1994). Judicious application of randomized designs. In K. J. Conrad (Ed), *New Directions in Program Evaluation*, *63*, 73-86.

Raudenbush, S. W. & Liu, Xiao-Feng (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.

Reichardt, C. S., & Rallis, S. F. (Eds., 1994). The qualitative-quantitative debate: New perspectives. *New Directions for Program Evaluation Monograph* (No. 61). San Francisco, CA: Jossey-Bass.

Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*, 33-38.

Rossi, P. H. (1994). The war between the quals and the quants: Is a lasting peace possible? In C. S. Reichardt and S. F. Rallis (Eds.). The qualitative-quantitative debate: New perspectives (pp. 23-36). *New Directions for Program Evaluation Monograph* (No. 61). San Francisco: Jossey-Bass.

Rossi, P. H. (1997). Advances in quantitative evaluation, 1987-1996. In D. Rog and D. Fournier (Eds.), *New Directions for Program Evaluation, 76*, 57-68. San Francisco: Jossey-Bass.

Schorr, L. (common-purpose.org, 2001).

Schorr, L. (1 December, 1999). Discussant. National Invitational Conference on Early Childhood Learning: Programs for a New Age. Alexandria, VA.

Schorr, L., & Yankelovich, D. (2000, February 16). What works to better society can't be easily measured. *Los Angeles Times*, B7.

Sechrest, L., Babcock, J., & Smith, B. (1993). An invitation to methodological pluralism. *Evaluation Practice*, *14*, 227-235.

Sechrest, L., West, S., Phillips, M., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments (pp. 15-35). In L. Sechrest (Ed.), *Evaluation Studies Review Annual*. Beverly Hills, CA: Sage.

Shadish, W. R. (1986). Planned critical multiplism: Some elaboration. *Behavioral Assessments*, *8*, 75-103.

Shonkoff, J. P., & Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early child development*. Washington, DC: National Academy Press.

Simpson, J. M., Klar, N., & Donner, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health, 85*(10), 1378-1386.

St. Pierre, R. (1983). Congressional input to program evaluation: Scope and effects. *Evaluation Review*, *4*, 411-436.

St. Pierre, R. G., Layzer, J. L., Goodson, B. D., & Bernstein, L. S. (1997, June). *National impact evaluation of the Comprehensive Child Development Program: Final report*. Cambridge, MA: Abt Associates.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.

Sue, S. (1999). Science, ethnicity, and bias. *American Psychologist*, *54*, 1070-1077.

Weiss, C. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. C. Kubisch, L. B. Schorr, and C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts*. Washington D. C.: The Aspen Institute.

Weiss, C. (1997). Theory-based evaluation: Past, present, and future. In D. J. Rog and D. Fournier (Eds.), *New Directions in Program Evaluation*, *76*, 41-56.

Weiss, H., & Halpern, R. (1991). *Community-based family support and education programs: Something old or something new*? New York: National Center for Children in Poverty, Columbia University.

Weissberg, R. P., & Kumpfer, K. L. (Eds.,2003). Prevention that works for children and youth. Special issue, *American Psychologist*, *58*, 425-490.

Yoshikawa, H., Rosman, E. A., & Hsueh, J. (2001). Variation in teen mothers' experiences of child care and other components of welfare reform: Selection processes and developmental consequences. *Child Development*, *72*, 299-317.

Social Policy Report is a quarterly publication of the Society for Research in Child Development. The *Report* provides a forum for scholarly reviews and discussions of developmental research and its implications for the policies affecting children. Copyright of the articles published in the *Report* is maintained by SRCD. Statements appearing in the *Report* are the views of the author(s) and do not imply endorsement by the Editors or by SRCD.

Electronic access to the *Social Policy Report* is available at the *Report*'s website:
http://www.srcd.org/spr.html

Subscriptions available at $20.00 to nonmembers of SRCD, single issues at $5.00, and multiple copies at reduced rates. Write SRCD Executive Office (srcd@umich.edu) or phone (734) 998-6578.

Purpose

Social Policy Report (ISSN 1075-7031) is published four times a year by the Society for Research in Child Development. Its purpose is twofold: (1) to provide policymakers with objective reviews of research findings on topics of current national interest, and (2) to inform the SRCD membership about current policy issues relating to children and about the state of relevant research.

Content

The *Report* provides a forum for scholarly reviews and discussions of developmental research and its implications for policies affecting children. The Society recognizes that few policy issues are noncontroversial, that authors may well have a "point of view," but the *Report* is not intended to be a vehicle for authors to advocate particular positions on issues. Presentations should be balanced, accurate, and inclusive. The publication nonetheless includes the disclaimer that the views expressed do not necessarily reflect those of the Society or the editors.

Procedures for Submission and Manuscript Preparation

Articles originate from a variety of sources. Some are solicited, but authors interested in submitting a manuscript are urged to propose timely topics to the editors. Manuscripts vary in length ranging from 20 to 30 pages of double-spaced text (approximately 8,000 to 14,000 words) plus references. Authors are asked to submit manuscripts electronically, if possible, but hard copy may be submitted with disk. Manuscripts should adhere to APA style and include text, references, and a brief biographical statement limited to the author's current position and special activities related to the topic. (See page 2, this issue, for the editors' email addresses.)

Three or four reviews are obtained from academic or policy specialists with relevant expertise and different perspectives. Authors then make revisions based on these reviews and the editors' queries, working closely with the editors to arrive at the final form for publication.

The Committee on Child Development, Public Policy, and Public Information, which founded the *Report,* serves as an advisory body to all activities related to its publication.