WP-02-19

Institute for Policy Research Working Paper

Randomized Experiments in Education: Why are they so rare?

Thomas D. Cook
Northwestern University

Thanks for feedback on an earlier draft are due to Tony Byrk, Lee Cronbach, Joseph Durlak, Christopher Jencks, and Paul Lingenfelter. The financial support of the MacArthur, Rockefeller and Spencer Foundations is gratefully acknowledged. Parts of this paper were read as the Jerry Lee Lecture at the University of Pennsylvania.

Introduction

Calls are heard to improve schools through innovations as diverse as school-based management, charter schools, vouchers, more effective teaching practices, higher standards, increased accountability, smaller schools, smaller classes, new technologies, and better trained teachers. Many “brand-name” reforms also exist, including Slavin’s Success for All, Levin’s Accelerated Schools, Comer’s School Development Program, Sizer’s Coalition of Essential Schools, Lezotte’s Effective Schools and Total Quality Management Schools. Claims also abound about other educational strategies like revisions to special education and bilingual programs, more phonics in early grades, constructivist learning that treats students as active learners, an end to social promotions, and better integration of schools with pre-schools, families and after-school activities. And these are only a subset of all the reform proposals recently made.

We later see that most of these ideas have never been seriously evaluated to learn how they affect student performance. This is surprising since there are many educational evaluators, and we might presume them to be interested in identifying “what works”. After all, Schools of Education include both self-styled evaluation specialists and substantive researchers who seek to specify factors that enhance student performance both to construct better theories and to impact on school practice. Academics outside of Education Schools also do educational evaluations, as do researchers in private firms that contract with Federal, state and local education authorities.

All these researchers have access to the generally preferred means for learning what works. Measuring individual student change is a long-recognized strength of educational research, and the randomized experiment is widely known to be the best tool for attributing observed student change to a particular policy option. Random assignment entails using the equivalent of a fair coin toss to create two or more initially equivalent groups. The option under consideration (“the treatment”) is then assigned to one group, while the other group is exposed to something else--often no explicit treatment but sometimes a qualitatively different one. If an experiment is properly maintained over time, any observed group differences at the end of a study cannot be due to differences in the average person in each group, for the assignment process renders such differences unlikely. Random assignment provides the best counterfactual describing what would have happened to treatment group students if they had not been exposed to the treatment (Rubin, 1974; Holland, 1986).

This theoretical rationale for random assignment is complemented by an empirical justification. When experimental results are contrasted with results from major design and statistical alternatives, different effect sizes are often found (e.g., Mosteller, Gilbert & McPeak, 1980; Lalonde, 1986; and Fraker & Maynard, 1987). Given the stronger logical warrant for random assignment, this suggests that non-randomized alternatives tend to provide more biased causal conclusions. Even when no differences are observed between experiments and quasi-experiments on the same topic, the variation in results across the quasi-experiments is greater than across the experiments (Lipsey & Wilson, 1993). This suggests that experiments are more efficient; they get to the right answer sooner. A pragmatic justification is also relevant. What if policy elites incorrectly concluded that Catholic schools are superior to public ones, and did something about this? What if they erroneously concluded that vouchers stimulate academic achievement, and did something about this? What if they falsely concluded that school desegregation does not affect minority achievement when it does, and acted accordingly? Incorrect causal conclusions have costs.

The superiority of random assignment for drawing inferences about the consequences of planned change attempts is routinely acknowledged in philosophy, medicine, public health, agriculture, statistics, micro-economics, psychology, criminology, prevention research, early childhood education, marketing and those parts of political science and sociology concerned with improving opinion surveys. It is also acknowledged in all the elementary education method textbooks we have consulted. However, we shall see that it is hardly practiced in education, especially for assessing the impact of educational interventions of obvious policy-relevance. Education is not unique in this paucity of experimental evidence. Random assignment is also rare in sociology, political science, macro-economics and management. Yet causal statements are routinely made in these fields, usually through linking substantive theory to qualitative or quantitative non-experimental analyses. This paper does not argue that correct causal conclusions come only from experiments. It does argue, though, that experiments provide a better warrant for such conclusions than any other method so that, if experiments can be conducted in schools, they should be. Not to use them requires a strong justification.

Over the last 30 years, self-ascribed educational evaluators like Alkin, Cronbach, Eisner, Fetterman, Fullan, Guba, House, Hubermann, Lincoln, Miles, Provus, Sanders, Schwandt, Stake, Stufflebeam and Worthen, have proposed many justifications for not doing experiments. Some of these educational evaluators want evaluation to provide individual school and district staff with immediately usable continuous feedback about strategic planning, program implementation, and student or staff performance monitoring. They want evalution to improve the organization and management of individual districts or schools, assuming that this will routinely improve student performance. Other educational evaluators want evaluation to contribute to developing general theories of the often complex constellation of forces that bring about some important effect. The aspiration is to identify generative processes that have effects over a broad set of circumstances, much like engaged time on task does as a cause of achievement. Neither of these purposes places any premium on directly observing student change and unambiguously attributing it to a single policy-related treatment.

This paper probes the validity of the intellectual arguments that educational evaluators have adduced for not doing experiments and for taking evaluation in directions away from identifying the effects of circumscribed causal agents of relevance to educational policy. Less emphasis is placed on the political and organizational factors within the Federal system of support for educational research that Vinoskis (2002) emphases in his explanation of the paucity of experiments. To the (unknown) extent that the justifications offered by self-ascribed educational evaluators also overlap with those of their more substantively oriented colleagues in Schools of Education, we also address the latter's objections. Although the specific reasons for downplaying experiments vary across educational evaluators, the total set of arguments can be divided into five types:

(1) <u>Philosophical arguments</u> designed to show that experiments (a) cannot provide unbiased tests of causal hypotheses, and (b) are predicated on a descriptive theory of causation that is less useful than explanatory theories of cause.

(2) <u>Practical arguments</u> asserting that experiments (a) can rarely be mounted in schools, and (b) when mounted, are often imperfectly realized because of

compromises to the planned treatment contrasts and the quality of individual treatment implementation.

(3) Arguments about undesirable trade-offs because experiments: (a) sacrifice external for internal validity; and (b) value causal conclusions so highly that a conservative bias results which overlooks useful findings indicated by more liberal criteria.

(4) Arguments that schools will not use experimental results because: (a) experiments meet the interests of Federal and State policy makers who are not major actors in educational policy; and (b) the experiment's logic recreates a rational decision-making model that does not describe how schools actually make decisions.

(5) Arguments that experiments are not necessary because better alternatives exist, including: (a) the intensive qualitative case studies preferred by self-styled educational evaluators; (b) the quasi-experiments conducted by substantive researchers who prefer design control over statistical control; and (c) the causal modeling preferred by substantive researchers who do longitudinal work in schools.

Any of the points above casts doubt on the wisdom or practicality of experimenting in schools. So, the number and variety of the arguments confers a genuine intellectual integrity, making it important to deal with each argument in turn, both to examine its validity and to assess its implications for creating a practical theory of school-based experimentation.

Philosophical Beliefs Adduced to Reject Random Assignment

a. Random assignment is epistemologically discredited. To philosophers of science, positivism connotes a rejection of realism, the formulation of theories in mathematical form, the primacy of prediction over explanation, and the belief that entities do not exist independently of their measurement. Although this epistemology has long been discredited, many educational researchers still use the term “positivism” but connote something less historically precise—viz., quantification and hypothesis-testing, both central to experimentation. Kuhn’s work (1970) is at the forefront of their reasons for rejecting positivist science. He argued that theories are "incommensurable"—that is, their postulates cannot be formulated as specifically as philosophical theories of verification or falsification require and are always subject to reinterpretation. He also argued that observations are inevitably "theory-laden"—that is, they are impregnated with researchers’ theories, hopes, wishes and expectations, thus undermining their neutrality for discriminating between truth claims. In refuting the possibility of totally explicit theories and totally neutral observations, Kuhn’s work seems to undermine science in general and experimentation in particular. Tending in the same direction are the views of other philosophers that educational evaluators like to cite, such as Lakatos, Harre and Feyerabend. Also relevant is their citation of descriptions showing how often scientists’ behavior in the laboratory deviates from the very scientific norms they espouse (e.g. Latour & Woolgar, 1979). All these sources are meant to indicate that science is an emperor without clothes.

However, the critique is overly simple. Even if observations are never theory-neutral, many observations have stubbornly reoccurred whatever the researcher’s predilections. As theories replace each other, most fact-like statements from the older theory are incorporated into the newer one, surviving the change in theoretical superstructure. So, even if there are no "facts" we can independently know to be certain, there are still many propositions with such a high degree of facticity that they can be confidently treated as though they were true. For practicing experimenters, the trick is to build multiple theories into how the data are collected, especially the perspectives of theoretical opponents. Independent replications are particularly important, therefore, provided they do not share bias in the same direction (Cook, 1985). Kuhn's work complicates what a "fact" means but does not deny that some claims to fact-like status are strong.

Kuhn is also correct that theoretical statements are never definitively tested (Quine, 1951; 1969). But this does not mean that individual experiments fail to probe theories and the causal hypotheses they generate. When an experiment produces negative results its advocates are not likely to accept the disappointing result. Instead, they invoke various methodological and substantive contingencies that might have changed the result--perhaps if a different outcome measure had been used or if a different population had been examined. Subsequent studies can then probe these contingency formulations. If the results again prove negative, this might lead to an even more complicated contingency hypothesis designed to explain the latest disconfirmation. A test of this revised hypothesis can then take place, and so on. After a time, this process runs out of steam, so particularistic are the contingencies that remain to be examined. The consensus seems to emerge that: "The program could not be shown to be effective under the many different conditions examined. Other conditions could still be probed. But they are so circumscribed that the reform will not be worth much even if it is effective under these conditions". Kuhn is correct that this process is social and not exclusively logical, and he is further correct that the predicament arises because program theory is not explicit enough to be definitively confirmed or rejected. But the reality of elastic theory does not mean that decisions about causal hypotheses are only social or that they are devoid of all empirical or logical content.

b. <u>Experiments are predicated on an overly simple theory of causation.</u> Randomized experiments test the impact of only a small subset of potential causes from within the world, often a single one. And at their most elegant, they can responsibly test only a modest number of interactions between treatments. So randomized experiments are best when a causal question is simple, sharply focused and easily justified. The theory of causation most relevant to this is variously called the manipulability, activity or recipe theory (Collingwood, 1940; Gasking, 1955; Whitbeck, 1977). It seeks to describe the consequences of a set of activities that can be listed as though they were recipe ingredients and can be actively manipulated as a whole to ascertain what effects the lumped manipulation has. The aim here is to describe the effects of a given cause.

However, the most esteemed theories of cause seek to ascertain the causes of a given effect. They want to explain rather than to describe "if-then" connections. One explanatory theory emphasizes identifying "generative processes" (Bhaskar, 1975; Harre, 1981). These are forces that bring about effects in a wide variety of circumstances, such as gravity as it affects falling, or a specific genetic defect as it induces phenylketonuria, or time on task as it facilitates learning. However, as simple as these examples seem to be they are replete with hidden causal contingencies. Thus, the genetic defect does not induce phenylketonuria if an appropriate diet is adopted early in life; and time on task does not induce learning if a student is disengaged or the curriculum meaningless. So, a second understanding of causation requires specifying all the contingencies (co-causes) that impact on an effect, including those that follow from a causal manipulation but are prior to the effect. Yet experiments were not designed for this purpose. They were designed to describe the effects of a multidimensional set of activities deliberately manipulated as a package. Experiments are only explanatory if the manipulations are chosen to help discriminate between competing theories; or if the processes mediating between a cause and effect are specified and measured; or if effect sizes vary in systematic ways across outcomes, populations or settings.

Cronbach and his colleagues (1980) maintain that explanatory theories of cause are more relevant to schools than the activity theory. They believe that multiple causal factors are implicated in all student or teacher change, and often in complex ways. Their model of real world causation is system-related, more akin to intersecting pretzels than to the experimenter's simple arrow from A to B. They rightly believe that no educational intervention fully explains an outcome; at most, it is just one more determinant of that outcome. Nor are effect sizes constant across student and teacher populations, or across school types and times. Causal contingency is the watchword, not simple generalization. Given this priority, the activity theory seems irrelevant since so few variables causally implicated in an effect can be simultaneously manipulated. Believing that experiments cannot faithfully represent a real world of multivariate, non-linear and often reciprocal causation, Cronbach and Snow (1976) searched for aptitude-treatment interactions--specifications that a treatment's effect depends on student or teacher characteristics. But they discovered few that were robust. This might reflect an ontological truth—Nature is not as contingently ordered as many theorists think. But it might also reflect the methodological problems associated with testing statistical interactions—e.g., under-specified theory, partially valid measures, imperfectly implemented treatments, truncated distributions, and non-interval scales. By itself, empirical research cannot inform us whether the real world is more or less contingently ordered than critics of the experiment contend.

Even so, the activity theory is clearly limited as a theory of causation. But to be limited is not to be useless. Notice how many useful conclusions about effective educational practices are today specified without qualification. For instance: Small schools are better than large ones; Time-on-task raises achievement; Summer school raises test scores; School desegregation hardly affects achievement; Assigning and grading homework raises achievement. Most educational researchers who espouse a highly contingent theory of causation nonetheless seem willing to accept some non-contingent causal statements. Critics of the experiment also seem to accept some minimally contingent statements—e.g., reducing class size increases achievement, provided that the amount of change is "sizable" and to a level under 20. Or, Catholic high schools increase graduation rates over public schools, but only in the inner-city. Commitment to an explanatory theory of causation has not stopped researchers from acting as though some educational change attempts result in dependable main effects or simple interactions.

Some causal contingencies are irrelevant to educational policy even if they are relevant to full explanation. For policy, the most important contingencies are those that, within normal ranges, modify the sign of a causal relationship and so help identify where a treatment is sometimes harmful. Less important are those interactions describing benefits that are generally positive even if they are more positive in some circumstances than others. Policy makers are often constrained in what they can do and cannot assign different treatments to different populations. But they are willing to advocate broad changes with differential effects provided that these effects are rarely negative. When policy concerns are paramount, it is possible to ignore many variables that genuinely contribute to fuller explanation.

We sympathize with those opponents of experimentation who believe that biased answers to big explanatory questions are more important than unbiased answers to smaller casual-descriptive questions. We also agree with them that random assignment depends on a less comprehensive and less esteemed theory of causation. But acknowledging these limitations does not undermine the justification for experiments. We still need to know about the effects of given causal agents. This is not a trivial knowledge need. And acknowledging the limitations of the activity theory of causation should embolden us to design future experiments with a greater explanatory yield than in the past. At a minimum, this means greater sensitivity to identifying moderator and mediating processes, and thus building into experiments the sampling and measurement particulars that such sensitivity requires. No more black box experiments.

Practical Reasons for Not Doing Randomized Experiments

a. Randomized experiments cannot be mounted. Education researchers were at the forefront of the flurry of social experimentation at the end of the 1960's and during the 1970's. Evaluations of Head Start (Cicirelli & Associates, 1969), Follow Through (Stebbins, St. Pierre, Proper, Anderson & Cerva, 1978), and Title 1 (Wargo, Tallmadge, Michaels, Lipe & Morris 1972) found few, if any, positive effects. These disappointing results engendered considerable dispute about methods, and many educational evaluators concluded that quantitative evaluation had been tried and failed. So, they turned to other methods. Other scholars responded by stressing the need to study school management and program implementation, believing them to be the reasons why results were so disappointing (Berman & McLaughlin, 1977; Elmore & McLaughlin,1983; Cohen & Garet,1975). However, the most criticized educational studies of the period did not involve random assignment. I know of only three randomized experiments on educational topics of policy relevance then available—studies of Sesame Street (Bogatz & Ball,1971), of the Perry Preschool Project (Schweinhart, Barnes & Weikart,1993), and of only 12 youngsters randomly assigned to a desegregated school (Zdep,1971). Since only the Zdep study took place in schools, it is not accurate to claim that policy-relevant randomized experiments had been tried in education and had failed. Indeed, to critique randomized experiments Cronbach et al (1980) had to re-analyze studies that had nothing to do with schools.

Nonetheless, many district officials do not like the focused inequities in school resources that random assignment generates, fearing negative reactions from parents and staff. They prefer it when individual schools choose the changes they will make, or when changes are district-wide. Principals and other school staff have similar preferences and have additional concerns about disrupting ongoing routines. Also, ethical concerns are often raised about withholding potentially helpful treatments; and some programs are meant to be universal under law, thus precluding the use of no-treatment control groups. Are experiments so unpopular and impractical that they cannot be used to study the effects of school improvement attempts?

It is manifestly false that experiments cannot be done in schools. We shall not consider here the small, controlled experiments done by education researchers with interests in cognitive science whose studies tend to be of little immediate policy relevance. On other topics there are many fewer experiments, but some. If we go to the policy areas specified in the first part of this paper's first paragraph, I could find no experiments on standards setting. The literature on effective schools reveals no experiments systemically varying the school practices that correlational studies suggest are effective. Recent studies of school-based management reveal only two randomized experiments, both on Comer's School Development Program (Cook, Habib, Phillips, Settersten, Shagle & Degirmencioglu,1999; Cook, Hunt & Murphy, 2000). There seem to be no experiments on Catholic or Accelerated or Total Quality Management schools. On vouchers there is a study by Witte (1998), sometimes re-analyzed (Greene, Peterson, Du, Boeger & Frazier,1996; Rouse,1998), plus a program of research by Peterson in several sites (Howell & Peterson, in press). On charter schools I know of no relevant experiments. On smaller class sizes, there are six experiments, the best known being the Tennessee class size study (Finn & Achilles,1990: Mosteller, Light & Sachs,1996). On smaller schools I know of only one randomized experiment, (Kemple, 2001). On teacher training I know of no relevant experiments. So, current knowledge of effectiveness depends heavily on methods less esteemed than random assignment.

Only two of the policy experiments cited above were conducted by researchers trained in Schools of Education or currently so affiliated. The best-known class size experiment was done by educators, but was popularized by statisticians and reanalyzed by an economist. The Milwaukee voucher experiment was done by political scientists and re-analyzed as a randomized experiment by political scientists and economists. The Comer studies were conducted by sociologists. The research on academies within high schools was done by an education-trained researcher working at Manpower Demonstration Research Corporation, a contract research firm with a strong economics background. The school choice work was done by political scientists.

To further illustrate the paucity of experiments, Nave, Meich and Mosteller (1999) report that not even 1% of the dissertations in education or of the studies archived in ERIC Abstracts involved random assignment. Casual reading of the major journals on school improvement (American Educational Research Journal and Educational Evaluation and Policy Analysis) tells a similar story. The National Institute of Child and Human Development's Congressionally mandated National Reading Panel reviewed nearly 2,000 published studies on the effects of phonemics (Ehri et al, 2001, a). Of these, 52 experiments and quasi-experiments met the criteria for inclusion. If the ratio of experimental and quasi-experimental comparisons in the meta-analysis is the same as the ratio of experimental and quasi-experimental studies, then about 23 of these 2,000 studies would have been randomized experiments--slightly more than 1%. In a related study of phonics (Ehri et al, 2001, b), 38 experiments and quasi-experiments were found, and of these 14 were randomized experiments. Unfortunately, the population of all studies claiming to be about the causal effects of phonics was not identified. Nonetheless, it is clear that experiments can be done on pedagogic topics. They are just not common.

Random assignment is much more common in school-based prevention studies designed to improve student health, to prevent school violence, or to reduce teen use of tobacco, drugs and alcohol (e.g., Cook, Anson & Wachli,1993; Peters & McMahon,1996; Durlak & Wells,1997a, b and 1998). Durlak's two reviews of prevention studies prior to 1991 including about 190 randomized experiments and 120 other studies. The number of experiments has certainly increased since then, given the rapid growth of prevention research. So, school-based experiments are common if the topic involves preventing negative behaviors. Experiments are also common in pre-school education. They were used with The Ypsilanti Perry Preschool Program (Schweinhart, Barnes & Weikart, 1993), The Abcedarian Project (Ramey & Campbell,1995 ; Campbell & Ramey,1991), the Comprehensive Child Development Program (Goodson, et al, 2000), Olds' home nurse visiting program (Olds et al, 1997), Early Head Start (Raikes & Love, 2002) and even the new Head Start study (R. Cook & Puma, 2002). Of course, pre-school experiments take place in homes or childcare centers rather than schools. But instruction is evaluated and changes in cognition and social behavior constitute the major outcomes, as in school-based pedagogic studies. So, school experiments are common if the topic is preventing negative behavior; and experiments are common in education at the pre-school level. But experiments on policy-relevant educational topics are rarer and, of those I know, most were conducted by contract researchers or academics outside of Schools of Education.

Why is there this disciplinary difference in the likelihood of conducting experiments? One possibility touches on subject matter. The prevention experiments tend to last less than a year; they do not involve changes in major school routines; they do not threaten the performance in math and language arts by which schools are held accountable; and the implementation is usually done by researchers rather than teachers. In contrast, pedagogical interventions are more multi-year; teachers are more often asked to deliver the treatment; changes are expected to teachers' established routines; and there is always the threat that performance in core competencies may not rise because of the intervention, compromising accountability goals. Arguing against this interpretation are two things. First is the personal report of Durlak that some of the190 prevention experiments he reviewed involve multi-year interventions, whole school changes and teachers delivering the treatment, though none speak to competencies in core academic areas. (However, it is not clear how many prevention studies combine all of the relevant features that tend to differentiate educational policy from prevention studies). And second, some pedagogic reforms do not require multi-year and whole school efforts. Yet even under the most propitious circumstances for experimentation, the low rate of pedagogic experiments documented earlier suggests they are rare.

A second explanation for the discipline-based difference in the frequency of experiments invokes political will and disciplinary culture. Random assignment is common in the health sciences because it is institutionally supported there by funding agencies, publishing outlets, graduate training programs, the clinical trials tradition, and practices in government health action agencies. Prevention studies in schools tap into a similar research culture, as does pre-school education where Congressional mandates play an ancillary role. Moreover, prevention and pre-school researchers tend to be trained in psychology, human development, public health and microeconomics--fields that value experimentation. Gueron (2002) emphasizes how important sponsor and researcher commitment are for getting experiments mounted

Contrast the norms and structures above with the situation in education. Reports from the Office of Educational Research and Improvement (OERI) are supposed to identify effective school practices. But neither the work of Vinovskis (1998) nor my own haphazard reading of OERI reports suggests any privilege accorded to random assignment. Moreover, one recent report I read on bilingual education repeated old saws about the impossibility of randomizing and claimed that alternatives are just as good--in this case poorly designed quasi-experiments. And at a recent foundation meeting on Teaching and Learning the representative of a reforming state governor spoke about a list of best practices being disseminated to all schools in his state. He did not care, and he believed that no governors care, about the technical quality of the research. His main concern was that there was consensus among education researchers about each practice. When asked how many of the recommended practices depended on evidence from randomized experiments, he guessed it would be none. Several nationally known educational researchers were also present, and they all agreed that such assignment probably played no role in generating the practices on the list. No one felt any distress at this. So long as such beliefs and feelings are widespread, there will never be the pan-support for experimentation in education that is currently found in health, agriculture, or health-in-schools. At present, there is much concern in the Office of Education to change this situation. Time will tell how much comes of this new priority, for government funders are only one source of professional norm-setting in education.

Principal ignorance of random assignment is not a likely cause of the low frequency of pedagogic experiments. When new programs are announced in schools, the demand for places sometimes exceeds the supply. In this situation, principals often resort to random assignment to determine who gets a place, being afraid of parental or staff reactions if they allocated slots by merit, need, " first come, first served" or teacher recommendation. Like other politicians, principals understand the benefits of random assignment when resources are limited.

What is needed to make randomized experiments more common in schools? There is no single road. In Cook et al (1999), random assignment was sponsored by the school district and all middle schools had to comply. Principals had no choice about participating or about the treatment they eventually received. The district took this step because a foundation-funded network of prestigious scholars--none from education--insisted on random assignment as a precondition for funding the program and its evaluation. In Cook, Hunt & Murphy (2000), the principal investigator insisted on random assignment as a precondition for collaborating with the program implementers. In Chicago, participation was restricted to schools where principals applied for the program and agreed in advance to live with the results of the randomization process. No principal had any difficulty appreciating the method's logic and most lived with its consequences for up to six years. (But not all. Of the 24 Chicago principals, one assigned to the control group dropped out immediately after the coin toss, and three of the treatment principals who retired at the end of the second study year had replacements who abandoned the program). Also important was honestly acknowledging up-front that the program might not be effective and promising the principals that, if they were assigned to the control condition, they would be the first to be offered the intervention at the study end, by when it might be improved. Schools were also paid for participating in the measurement process, and a year was set aside for recruitment.

Contrast this with what happened very recently when the developer of one of the nation's best-known school reform packages tried to evaluate his program using random assignment. He used letters and e-mail to solicit schools to volunteer to be in the experiment. His staff then followed up and found few schools willing to be assigned once the random assignment was explained to them. Nonetheless, the federal funders continued to insist that a randomized experiment be done. So, the program designer developed a within-school experiment in which some grades get the intervention but others do not, even though cross-grade contamination might occur in this circumstance. Given the facts above, can one seriously imagine the developer's staff informing schools they were being recruited because it was not clear that the program would work! Can we imagine them asserting that prior (quasi-experimental) research on the program's effectiveness was not definitive when this same research had earlier been used to argue for continued program funding? Developers are, and should be, passionate advocates for their programs, not brokers of honest appraisal. And how much does a program developer in education know about practical ways to implement random assignment if, as in this case, he has never done such a study before and if education as a field has little recent practical history with such studies? In the policy realm, random assignment should be in independent hands and carried out by staff with a recent history of successful randomization in complex field settings.

Which conditions are most conducive to random assignment? Applying the principles in Shadish, Cook & Campbell (2002) to schools suggests that such assignment is most feasible when: treatments are shorter; teachers training is not required; patterns of coordination among school staff are not modified much; the demand for an educational change outstrips the supply; different treatments with similar goals are compared; the units receiving different treatments cannot communicate with each other; and when students are the unit of assignment rather than classrooms or classrooms rather than schools. Thus, it should be easiest to study different curricula at random; to introduce new technologies at random; to give students tuition rebates for Catholic schools at random; to assign homework variants at random; to assign teachers trained in different ways at random. Such studies will not be easy. But all should be possible, given researcher will to make random assignment happen and researcher knowledge about how to make it happen.

b. <u>Even when mounted, many planned between-treatment contrasts are compromised</u>. Random assignment creates treatment group equivalence at the pretest. But it is at the posttest that groups should be equivalent in everything except treatment exposure. Sometimes, different kinds of students drop out of the various treatment groups, creating a subsequent group non-equivalence that threatens the integrity of an experiment. Such differential attrition is most likely when treatments vary in intrinsic value, and since many independent variables involve deliberately created resource differences, in policy studies differential attrition is not a remote possibility. However, it can be routinely minimized, if not always completely prevented. The key is eliciting a strong initial staff commitment to staying in the study, providing modest payments to the units experiencing less desirable treatments, and closely monitoring treatment implementation, including monitoring to detect and deal with early drop-out trends.

With long-lasting treatments, an additional difficulty arises. Officials leave schools, and their replacements sometimes want to jettison their predecessor's innovations and introduce their own, especially when the predecessor's reform initially disrupted school routines and so has not yet generated a strong core of supportive teachers. Little can be done in this situation which, in my experience, often occurs in the first years of whole school reform. If possible, the schools lost to intervention should remain within the measurement framework. But this is not always possible. And such strategies are not perfect. Despite taking the precautions above, Cook, Hunt & Murphy (2000) still lost four of their 24 schools. (Fortunately, they built-in strong fall-back quasi-experimental options, pretest values on the outcome variable being a minimum for this). But even with some differential attrition, the resulting bias is likely to be less than the bias due to school or teacher self-selection from the start; and statistical selection controls are better the smaller the initial bias and the better selection has been directly observed (Holland, 1986).

Experiments can be compromised by treatment-crossovers as well as differential attrition. These cross-overs occur when units in one treatment condition experience intervention particulars destined for another, thereby reducing the size of the treatment contrast and increasing the chances of falsely concluding there is no treatment effect. One way cross-treatment borrowing occurs is when the units receiving different treatments can communicate with each other. To circumvent this, researchers should work with physically separated units. This is often easy, though not always. However, the more schools are separated the higher the research budget becomes and the more schools are needed to meet sample size requirements.

Extensive treatment cross-overs may be rare, though. Cook et al (1999) documented that only three of ten control schools borrowed any program elements, and none borrowed the program's most central elements. They did not have access either to School Development Program facilitators or to treatment-specific professional development opportunities, including trips to the program developers at Yale. Of the three documented cross-overs, one occurred because a treatment principal was married to a teacher from a control school and they talked about the intervention. Another was caused by the daughter of a Yale program official teaching in a control school and inviting her father to give some lectures at the school. The third involved a control principal becoming interested in the program, reading up on it, and trying to implement some of its practices without formal program support. Also relevant is that the district program coordinator also did some district-wide professional development and some program details entered into what she taught. So, cross-treatment borrowing occurred, but it was not universal across schools, it involved some but not all program details, and the borrowed particulars varied from school to school. None of this would have been detected, of course, without sensitivity to the possibility of treatment cross-overs and without collecting annual data on the matter.

The eternal hope is that treatments will be so innovative that the control units will experience none of the treatment particulars. But even without direct communication control units often have experiences that overlap with those planned for the treatment group. It is as though program designers are inspired by ideas that appear new but that are, in reality, only a little ahead of the emergent professional mainstream. During the study, they then enter into that mainstream and so controls pick them up. For evaluation to be maximally useful, it may be that program designers need to be more original. But even when the various treatment groups have become closer in content, this does not mean that the planned contrast is useless for policy purposes. It is still worth learning whether the planned treatment adds something over and above the newly emerged status quo. Program developers dislike this since they believe that their program can be better than the former status quo without being better than the new one. Particularly galling for them is the possibility that their own ideas might have co-caused the new status quo, thus heightening the bar over which their own program now has to jump. Developers want to learn about the effects of their planned treatment at its maximal point of contrast, and so they prefer it be evaluated against the best approximation to the total absence of anything resembling their treatment.

The usual approximation to this requires measuring treatment fidelity on each unit in each treatment condition and then using this fidelity measure as the "independent variable" in analyses of the outcome. A selection problem obviously results. However, if such stratification occurs within a randomized experiment, this is one of the few situations where instrumental variables credibly deal with selection (Angrist, Imbrens & Rubin, 1996), permitting two distinct and important questions to be answered. First, is there an effect of the original treatment assignment without regard to variation in treatment exposure within or between groups—the so-called "intent to treat analysis"? Second, is there an effect of the "treatment" students actually received, irrespective of their original treatment assignment? Is it not ironic that the currently most defensible test of the causal impact of self-selection occurs if this self-selection takes place within the framework of a randomized experiment?

c. <u>Random assignment assumes fixed theory and standard implementation, but neither of these treatment-specific assumptions is valid in schools.</u> Experimental results are easier to interpret when the intervention is the product of strong substantive theory, when the achieved implementation faithfully reflects treatment-specific program theory, and when the within-treatment variation in implementation is minimal. These conditions are not often met. Schools tend to be large, complex social organizations characterized by multiple simultaneously occurring programs, disputatious building politics and conflicting stakeholder goals. Management is all too often weak and removed from classroom practice, and day to day politics can swamp effective program planning and monitoring. So, many reform initiatives are implemented highly variably. Indeed, when different educational models are contrasted in the same study, the between-model variation is usually small relative to the variation between schools implementing the same model (Rivlin & Timpane, 1975; Stebbins et al, 1978). In school research, it is not realistic to assume standard program implementation or total fidelity to program theory (Berman & McLaughlin, 1977). To those who assume that schools have severe management and implementation problems experiments must seem premature.

However, educational research does not need to assume such complex organization. An earlier model treated schools as physical structures with many self-contained classrooms in which teachers tried to deliver effective curricula using instructional practices that had been "shown" to enhance student performance. This approach privileged curriculum design and instructional practice over the school-wide factors that now dominate—e.g., strong leadership, a building-wide communitarian climate, a focus on learning, undertaking multiple forms of professional development, and creating supportive links to the outside world. Many important consequences follow from how schools were re-conceptualized. One is the lesser profile accorded to curriculum and instructional practice and to what happens once the classroom door is closed. Another is the view that random assignment is premature, given the presumption that its implementation depends on positive school management and quality program implementation. And another is the consequence that quantitative techniques are of lesser value, since school management and culture are best understood through ethnographic case studies.

Advocates of random assignment will not be credible if they assume treatment homogeneity or setting invariance in educational contexts. However, random assignment does not require well-specified program theories, good management, standard implementation or treatments that are totally faithful to program theory. Experiments primarily protect against bias in causal estimates; and only secondarily against imprecision in these estimates. So, the complexity and heterogeneity of schools leads to the need for larger school sample sizes and the need to anticipate and measure specific sources of variation in order o reduce their unwanted influence through statistical control. But just as importantly, implementation quality should be studied as a dependent variable to ascertain which types of schools and teachers implement the program better. Variable implementation is important in its own right, as well as having implications for budgets and sample sizes. We should also not forget that few educational interventions will be standardized once they are implemented as formal policy. So, why standardize them in an experiment? Treatment standardization is desirable for researchers interested in testing the substantive theory behind a treatment and for those interested in assessing an intervention's potential as policy. But it is not desirable for those seeking to determine a treatment's likely effects in settings where standard implementation cannot be expected.

Random assignment entails undesirable trade-offs

a. <u>Increasing internal validity decreases external validity</u>. Random assignment prioritizes on unbiased answers to descriptive causal questions. But few educational evaluators share this priority and most believe that it compromises more important research goals. Cronbach (1982) rejects the assertion that internal validity is the sine qua non of experimentation (Campbell & Stanley, 1963) because of the neglect this implies for external validity. Experiments are clearly limited in time and space, and nation-wide experiments are very rare. Most experiments are limited to the sub-class of schools willing to surrender choice over the treatment they will receive and to tolerate the in-school measurement of implementation, mediating processes and individual outcomes. What kinds of schools are these? Cronbach prefers to sample from a more representative population of schools even if less certain causal inferences result from this.

Science values results that are general. This includes the discovery and explanation of generative causal processes like gravity, relativity, DNA, nuclear fusion, aspirin, personal identity, infant attachment, or engaged time on task. Constructs like these index instantiating processes capable of bringing about effects in many different contexts and so are more general than learning whether one form of instruction affected achievement at a particular time in the particular sample of schools that happened to volunteer for an experiment. Many educational evaluators want their field to attain general understanding through explaining why programs work. Thus, they are prepared to tolerate more uncertainty than most scientists about whether a program does in fact work—especially if learning that they do work depends on experiments whose circumstances do not closely approximate educational practice. Educational evaluators espouse the traditional scholarly goal of full explanation, but they reject the quantitative methods usually preferred for this. Cronbach has even asserted that the methods of the historian, journalist and ethnographer suffice for learning about what happened in an educational reform and why.

An obvious problem with this high priority assigned to external validity is that it has not led education evaluators to reliably learn what works. It is now 30 years since vouchers were proposed, and we still have no clear answers about them. It is 30 years since Comer began his work on the School Development Program, and almost the same situation holds. It is 15 years since Levin began accelerated schools, and here too we have no experiments and no answers. The Obie-Porter legislation cites Comer's program as proven effective. But when the legislation passed, the relevant evidence consisted of testimony, studies by the program's own staff that used primitive quasi-experimental designs, and one interrupted time-series study that confounded the court-ordered introduction of the program with a simultaneously ordered reduction of about 40% in class sizes (Comer, 1988). Of the other whole school programs, only Success for All has been evaluated moderately well (for a summary see Herman, 1998). But even here, the evaluations have not been independent of the developer, the treatment assignment has never been random, and in no school did fewer than 80% of the teachers agree to the program. When trade-offs are made that favor generalization over cause, this risks ending up with the current state of affairs. Many studies exist in many districts at many times, but none is worth much as a study of cause. Experiments are not meant to be representative; they test causal claims.

But causal statements are more useful if they come from experiments where the sampling particulars permit tests of generalization across various types of students, teachers, settings and times. The hope is either to demonstrate empirical robustness or to identify the boundary conditions under which an effect occurs. The two keys here are research questions that are crystal clear about the populations targeted and then the use of sampling procedures that represent these targets and make heterogeneous all the other irrelevancies that might limit generalization. So, formal sampling is one way to increase external validity within experiments. The ideal is random selection followed by random assignment so as to achieve an unbiased causal estimate that generalizes without bias to a pre-specified population of schools, students, teachers or sites.

As admirable as this is, there are many reasons why sampling units with known probability has rarely been the path to generalization in the experimental sciences. Volunteering to be in a study is usually required, and this limits generalization. And many of the populations it is practical to sample without volunteering are of only parochial interest. Moreover, random sampling is hardly relevant to estimating causal relationships that might vary by historical period, and it cannot be used to select the outcome measures and treatment variants that are used to represent general cause and effect constructs. So, bench scientists use a different generalization model, one that emphasizes how consistently a causal relationship replicates across multiple sources of heterogeneity (Cook, 1993). The operative question is this. Can the same causal relationship be observed across different laboratories, time periods, regions of the country, and ways of operationalizing the cause and effect? This heterogeneity-of-replication model underlies current practice in both clinical trials and meta-analysis and it permits purposive rather than random sampling to be used. Vital is only a heterogeneous sampling plan with respect to people, settings, operational definitions and times--though multi-site clinical trials typically sample only one time period.

Heterogeneous, purposive sampling is not an easy path to follow for increasing external validity while maintaining high internal validity, especially in education which has no tradition of multi-site clinical trials with national reach. More typical are individual school experiments with unclear reach, being done only in Milwaukee or Washington or Chicago or Tennessee. In addition, few reform efforts in education have a fixed protocol. So, we can implement vouchers, charter schools or Total Quality Management in many different ways across districts and even within them. Indeed, the Comer programs in Prince George’s County, Chicago and Detroit are different from each other in many ways, given how much latitude districts are supposed to have in how they define and implement that program's specific details. This means that many educational treatments will require even larger samples of settings than do clinical trials in medicine where the between-site variation in protocols is almost certainly less than in schools.

Very large experiments may not be wise, however. The physical sciences have progressed in generalizing because knowledge claims are routinely replicated in the next stage of research on a phenomenon. But in research areas with weaker (and more expensive) traditions of replication--as in education--replication cannot be so haphazard. We need experience-filled theories of replication. Should replication depend on conducting a number of smaller experiments staggered over the years, each adequately statistically powered? Does it make sense to conduct even more experiments, but smaller ones, many of which are inadequately powered—as seems to be the case in school-based prevention studies? Is meta-analysis the only serious answer to the causal generalization problem, and so patience becomes a needed policy virtue because of the time needed to build up a data base? Whatever the merits of particular forms of phased programs of experiments, the point is that individual experiments vary in their sampling reach and in their connections to solid findings from the past. Single experiments rarely produce definitive answers, however large they are. And they certainly do not answer all ancillary questions about the contingencies on which a causal relationship depends.

We have seen that causal generalization can be understood as a single causal estimate for a given population (as in the formal sampling tradition) or as an average effect size derived from heterogeneous studies of the same hypothesis (as in synthesis methods). But causal generalization can also be understood as identifying generative causal processes. For instance, engaged time-on-task is presumed to stimulate achievement through activities as diverse as more homework, summer classes, longer school days, and more interesting curricula--procedures that can be implemented in any school district in any country at any time. The methods for identifying such explanatory processes place relatively little weight on sampling, instead requiring the collection of data about each of the variables in the presumed generative theory or getting historians and ethnographers to explore what happened and why in each treatment group. Fortunately, it is easier to build these explanatory methods into individual experiments than it is to sample at random or to add populations to the sampling design. In whatever ways are feasible, experiments should be designed to explain the consequences of interventions and not just to describe them. This means adding to an experiment's measurement and sampling plans and abjuring black box experiments.

b. Prioritizing scientific purity over utility. Critics contend that experimenters value uncertainty reduction about cause so much that conservative criteria are used to protect against wrong inferences, with the result that many effective programs are judged to be ineffective. One example of this is use of the traditional statistical criterion of $p <.05$ rather than, say.25. In deciding whether to adopt a potentially life-saving therapy for a loved one, would experimenters not use the more liberal risk calculus? Why be different in science? Should statistical traditions be so strict that schools not implementing the treatment are included in the analysis as though they had been treated? What about purist experimenters who refuse to explore the data for unplanned treatment interactions with student or teacher characteristics, or who view unplanned variation in implementation as a cause for concern rather than an opportunity to explore the origins and consequences of this variation? And why should one persist with the original research question if a more useful question has emerged during a study, even if unbiased answers to this new question are not possible? Better relevant than pure is the implication of these critical questions—all the more so since many experiments take so long to plan, mount and analyze that the answers they provide turn out to be of more antiquarian than contemporary interest.

Experiments do tend to be so preoccupied with bias protection that other types of knowledge are secondary. But they need not be so secondary. There is no compelling need for stringent alpha rates; only statistical convention is at play here, not statistical wisdom. Nor need one restrict data analyses to the intent-to-treat group, though such analyses need to be reported. Nor need one ignore all statistical interactions, though probing them should be done with substantive theory and statistical power in mind, and conclusions about substantive interactions should be couched more tentatively than conclusions about main effects. Researchers can also try to replicate experimental results generated from limited samples by re-analyzing data about similar constructs collected from formally representative samples and subjecting them to the best available non-experimental analyses. Finally, many controlled experiments will be improved by collecting ethnographic data in all the treatment groups in order to help identify mediating processes and unintended outcomes, also providing continuous feedback to the treatment and control schools alike. Experiments need not be as rigid as many clinical trial texts paint them.

Experimental Results are not likely to be used in Educational Policy

<u>a. Experiments prioritize on the information needs of central decision makers who are not important in the decentralized American educational system</u>. Most of the funds spent on education come from local sources, next most from states, and least from Federal sources. Yet experiments are least often designed to help local school staff. Indeed, staff members rarely seem to want what an experiment produces--information summarizing what a reform has achieved across a sample of schools. They want

information about their own school only, and they want it when they need it, not just at the end of a study. Putting the needs of local service deliverers above those of amorphous state and especially Federal policy-makers prioritizes on utility over truth and on continuous feedback over final reports. A letter to the New York Times captures this: "Alan Krueger ... claims to eschew value judgments and wants to approach issues (about educational reform) empirically. Yet his insistence on postponing changes in education policy until studies by researchers approach certainty is itself a value judgment in favor of the status quo. In view of the tragic state of affairs in parts of public education, his judgment is a most questionable one." (W. M. Petersen, April 20, 1999).

It is a mistake to believe central decision-makers are powerless in education. At both the Federal and state levels their role is steadily increasing. They are especially powerful in inner cities where the proportion of dollars from local taxes is lower than in the suburbs. They are also especially powerful in certain programs like special education and bilingual education. Moreover, Congress and state government are important sources of political pressure to improve schools in the belief that improved human capital will keep the economy strong in a global context where low wage industries have moved abroad. It is also a mistake to believe that experiments necessarily preclude continuous feedback to schools. The key requirements are only that such feedback be provided similarly in all treatment conditions and that no experimental contrast results be presented. But providing continuous feedback should not be done willy-nilly, for the information provided becomes part of the study context and limits generalization to settings where feedback is part of program design.

It may be that local school personnel tend to use information from site-specific management studies more than from experiments, though this is not clear yet. But experiments are more likely to get into the corpus of findings in textbooks and training manuals, thus influencing the next generation of teachers and principals. The anticipated benefit from experiments is their potential reach across a nation, either through results getting into textbooks or policy decisions. It is a mistake to limit understanding of research use to immediate use by the districts or schools in a study. Nonetheless, experimenters should do all they can to generate research questions that have an obvious long-term utility to a wide range of policy actors, including local school personnel.

b. Experimentation recreates a classical model of rational decision-making that has been disproved. Theories of rational decision-making require analysts first to lay out the alternatives (the treatments). Then one decides on decision criteria (the outcomes). Next, one collects information on each criterion for each treatment (data collection). And finally, one uses the observed effect sizes and whatever utilities can be attached to them to make a decision about the merits of the contending alternatives. Empirical work on how social science data are used in policy reveals that such use (termed instrumental use) is rare (Weiss & Bucuvalas, 1977; Weiss, 1988). Instead, use is more diffuse and better described by an "enlightenment" model that involves information blended from existing theories, personal testimony, extrapolations from surveys, the consensus of a field, empirical claims from experts who may or may not have interests to defend, and novel concepts that are au courant and broadly applied--like social capital in sociology today. Describing research utilization in this enlightenment fashion extends no special privilege to science in general or to experiments in particular.

Empirical research also notes that use decisions are multiply rather than singly determined, with central roles being played by politics, personality, windows of opportunity and values. Also, many decisions are not made in a systematic sense, but are rather slipped into or accrete, with earlier small decisions constraining later larger ones. In addition, official decision-making bodies change in personnel, with new persons and issues replacing older ones. When studies take time to complete--as with most experiments—the results may not be available until the policy agenda has already changed. Research use is much more complex than simply making an evidence-based rational choice

Critics also note that experiments rarely provide uncontested verdicts. Disputes typically arise about how well the original causal question was framed, about whether the claimed results are valid, about whether all relevant outcomes were assessed, and about whether the proffered recommendations follow from the results. The logical control over selection that makes experiments so valuable does not mean that all quibbles about causal claims are put to rest. Consider the very different conclusions offered about whether and where effects are warranted in the Milwaukee voucher study (Witte, 1998; Green, Peterson, Du, Boeger & Frazier, 1996). Consider, also, the different effect sizes generated from the Tennessee class size experiment and (Finn & Achilles, 1990; Mosteller, Light & Sachs, 1996; Hanushek, 1999). Sometimes, real scholarly disagreements are at issue, while in other cases the disputes reflect stakeholders protecting their interests. Policy insiders use multiple criteria for making decisions, and scientific knowledge of causal influences is never uniquely determinative.

Close examination of claims that policy changes were made <u>because of</u> experimental results suggests some oversimplification. The Tennessee class size results are consonant with the results of an earlier meta-analysis (Glass & Smith, 1979) and with theories that say children gain more if they are engaged and on-task. The results also conform with teachers' hopes and expectations, as well as with parents' commonsense notions. Moreover, when the results were delivered the Tennessee governor had national political ambitions, thought he could increase educational investment, and he knew his actions would be popular with both teacher unions and business interests. So, any policy change he might have made could not be attributed to random assignment alone. However, Tennessee did not reduce class sizes, due to financial cost and lack of teachers at this time of national teacher shortage (Boruch, 2000). Other states did make the change to smaller classes, though; and to do so California had to recruit teachers from other (less wealthy) states, possibly exacerbating state-level inequalities. There was also some teacher-poaching across district lines within California, favoring the wealthier districts. Also, new teachers were recruited from corporate venues, assuming they would be effective despite scant formal training and no experience as teachers. And smaller classes meant that some California students had to be located in sub-optimal buildings, since lower class sizes typically require more space. Experiments like the Tennessee study exist on a smaller scale than would typically pertain if the services they test were to be implemented state- or nation-wide. This scale issue is serious since program dynamics can be different on the larger scale and thus entail a different pattern of effects than achieved on the smaller scale (Elmore, 1996).

There is some substance to the idea that the theory of use buttressing randomized experiments is at odds with the ways social science data are used. But the objections are exaggerated. Instrumental use does occur (Chelimski, 1987), and more often than the very low base rates implied in most research denigrating instrumental usage. Moreover, some results are probably more widely disseminated because random assignment confers credibility on them in many quarters. This happened with the Tennessee class size study and the pre-school studies cited earlier. Studies can even be important if political events have rendered the results obsolete when they are announced. This is because many policy initiatives are recycled later--as with vouchers—and because the texts used to train professionals in a field often describe past studies that throw light on particulars of professional practice (Leviton & Cook, 1983).

There is also no necessary trade-off between instrumental and enlightenment usage. Experiments also contribute to enlightenment. They teach us about the kinds of interventions that can be better or worse implemented and about how principal turnover seems to affect school management. They inform about the low utility of theories that fail to specify what happens once the teacher closes the classroom door, about the kinds of principals who are most attracted to school-based management, and about the kinds of teachers most open to professional development. The era of black box experiments is long past. We now want to learn, <u>within experiments</u>, about the determinants and consequences of implementation quality and about the viability of the substantive theory under-girding program design. We also want to collect qualitative and quantitative data so long as the data collection protocol is identical in all treatment groups. And we want to get all stakeholder groups involved in formulating experimental questions and in interpreting the relevance of findings. These steps help generate enlightenment and thus make the experiment more like what its critics claim it is not.

Random assignment is not needed since better alternatives already exist.

<u>a.Intensive Case Studies</u>. No method will die, whatever its imperfections, unless a demonstrably better or simpler method can replace it. Educational evaluators believe that superior alternatives to the experiment already exist. The alternative they generally prefer is the intensive case study. The main reason for this is its flexibility of purpose. It can be used to appraise the theory of the program, to assess implementation quality, to record program re-design, to identify whether planned outcomes occur, to identify

unplanned effects, to estimate sub-group effect differences, to probe reasons for the effects claimed, and to assess the relevance of the findings for different stakeholder groups. By itself' the experiment cannot match this flexibility, having been designed only to answer one of the questions above.

Few advocates contend that case studies reduce as much uncertainty about cause as an experiment. But they do assert that case studies can reduce such uncertainty to an acceptable level. Laypersons often believe the clams of investigative journalists, historians and ethnographers, none of whom do not experiments. When done carefully, their methods have a methodological base not much different from quantitative science. That is, observations are first used to develop a hypothesis about what works. Researchers then think through other implications of this hypothesis and collect data relevant to these implications. This round of observations is then used to revise the last version of the hypothesis, and so on until closure is reached. This is basically an empirical, falsificationist hypothesis-testing procedure, and theorists of ethnography have long advocated it (e.g., Becker, 1958). Its results should be particularly rich for explaining why findings came about because ethnography demands close attention to processes as they unfold at different stages in a program's progress. Hard thought and non-experimental empiricism can surely reduce some uncertainty about cause--sometimes even all the uncertainty, though it will usually be very difficult to know when this last happens.

Nonetheless, intensive, qualitative case studies do not generally reduce as much causal uncertainty as experiments. The case studies rarely involve a convincing causal counterfactual. The absence of comparison groups makes it difficult to know how a treatment group would have changed in the absence of the reform under analysis. Adding control groups helps, but unless they are randomly created it will not be clear whether the two groups would have changed at similar rates over time. Whether intensive case methods reduce enough uncertainty about cause to be generally useful is such a poorly specified proposition that we cannot answer it. Still, it forces advocates to note yet again that experiments are best justified when a high standard of uncertainty reduction is required about a manifestly important causal claim.

Factors other than flexibility also favor case studies. First, schools are probably less squeamish about collaborating with ethnographers than experimentalists. Second, case studies produce human stories that can be used to communicate the results. And third, feeding interim results back to the teachers and principals with whom an ethnographer has ongoing relationships may be especially likely to generate local use of the data. This is less grandiose than affecting large numbers of districts and schools, but educational evaluators like Stake, Guba and Lincoln consider local use as the ultimate desideratum because they doubt that schools will comply with policy dictates from outside. Intensive case studies have many advantages, and we value them.

However, we value them most for their role within experiments rather than as alternatives to them. They complement an experiment whenever a causal question is central but it is not clear how successful program implementation will be, why implementation shortfalls may occur, what unexpected effects are likely to emerge, how respondents interpret the questions asked of them, what the casual mediating processes are, etc. Since these questions are important and not relevant to experimental functions per se, qualitative methods have a central role to play as adjuncts within experimental work on educational interventions. They cannot be afterthoughts.

b. Quasi-Experiments. Most researchers who do educational evaluations do not think of themselves as evaluators. They are primarily substantive researchers who want to test the effectiveness of changes within their own sub-field. They mostly use quasi-experiments, as in the early evaluations of Comer’s program noted earlier and the studies Herman (1998) detailed for Success for All and other whole school reform ideas. According to Herman, qualitative studies are also rare. This last is surprising, given how much emphasis theorists of educational evaluation have placed on qualitative work. Has their advocacy had little influence on their own colleagues? We do not know. But it is not easy to integrate outcome-based qualitative and quantitative case studies into a summary picture of the effects of a school reform.

Quasi-experiments are identical to experiments in purpose and in most structural details, the defining difference being no random assignment. Quasi-experiments use design rather than statistical controls to create the best possible approximation (or approximations) to the missing counterfactual that random assignment would have generated. These design controls include matched comparison groups, age or sibling controls, pretest measures at several times before a treatment begins, interrupted time-series, assigning units based solely on a quantitative criterion, assigning the same treatment to different groups at different times, and building multiple outcome variables into studies, some of which should theoretically be influenced by a treatment and others not (Corrin & Cook, 1998; Shadish, Cook & Campbell, 2002). Quasi-experimental designs are created through a mixing process that tailors the research problem and the resources available to the best design that can be achieved by mixing the design elements above.

In some quarters, "quasi-experiment" has been used promiscuously to connote any study that is not a true experiment, that seeks to test a causal hypothesis, and that has any type of non-equivalent control group or pre-treatment observation. Yet Campbell and Stanley (1963) and Cook and Campbell (1979) labeled some such studies as "generally causally uninterpretable", and many of the studies that educational researchers call "quasi-experiments" are of this last kind. They lag far behind the state of the art. Reading quasi-experimental studies of educational reform projects is dispiriting, so weak are the designs and so primitive are the statistical analyses. All quasi-experimental designs and analyses are not equal. Recent advances in the design and analysis of quasi-experiments are not getting into educational research where they are needed.

The best estimate of any quasi-experiment's internal validity is to compare its results with those from a randomized experiment on the same topic. When this is done systematically, quasi-experiments are more likely to be biased and less efficient than randomized experiments. So, in areas like education where few studies exist on most of the reform ideas being currently debated, randomized experiments are particularly needed. It will take fewer of them to arrive at what might be the same answer that better quasi-experiments would eventually achieve. Moreover, nearly all scholars trust the answers from experiments more.

c. <u>Theory-Based Evaluations</u>. It is currently fashionable in many foundations and some scholarly circles to espouse a non-experimental theory of evaluation for use in complex social settings like communities and schools (Connell et al, 1995). The method depends on explicating the substantive theory behind a reform initiative and detailing all the flow-through relationships that should occur if the intended intervention is to impact on a major distal outcome like achievement gains. The method also requires measuring each construct in the substantive theory and then analyzing to assess whether the postulated relationships have actually occurred in the data in the predicted time sequence. With shorter time periods, the data analysis will involve only the first part of the postulated causal chain; but over longer periods the complete model might be testable. This conception of evaluation places a primacy on very specific theory, high quality measurement, and the valid analysis of multivariate explanatory processes as they unfold in time.

The claim is that such theory-based evaluation can function as an alternative to random assignment that is more feasible. First, it does not require a causal counterfactual constructed through random assignment or matched comparison groups. Only the group experiencing a treatment is needed. Second, obtaining data patterns congruent with program theory is assumed to validate that theory. This is an epistemology that does not require explicitly rejecting alternative explanations, merely demonstrating a close match between the predicted and obtained data. Finally, the theory approach does not depend on attaining the end points typically specified in educational research--usually a cause-effect relationship involving student performance. Instead, when the initial phases of the program theory are corroborated, this can be used to argue for maintaining the program because it might be effective if data on more distal criteria were collected. Such corroboration also defends against prematurely concluding that a program is ineffective when insufficient time has elapsed for all the intervening processes to occur that might bring about the ultimate change. So, in this theory-based approach to evaluation no requirement exists to measure the endpoints most often valued in educational studies.

Few advocates of experimentation will argue against the greater use of substantive theory in evaluation. Such measurement will make it possible to probe, first, whether the intervention led to changes in the theoretically specified intervening processes and, second, whether these processes could then have plausibly caused changes in distal outcomes. The first of these tests will be unbiased because it relates each step in the causal model to the planned treatment contrast. But the second test will entail selection bias if it depends only on stratifying units according to the extent the postulated theoretical processes have occurred. Still, quasi-experimental analyses of the second stage are worth doing provided that their results are clearly labeled as more tentative than the results of planned experimental contrasts. The utility of analyzing theoretical intervening processes in experiments is beyond dispute; the issue is whether such measurement and analysis can alone provide an alternative to random assignment.

There are reasons for skepticism (Cook, 2000). First, it has been my experience writing papers on the theory behind a program with its developer (Anson, Cook, Habib, Grady, Haynes & Comer, 1991) that the theory is not always very explicit and could be made explicit in several different ways, not just one. Is there a single theory of a program, or several possible versions of it? Second there is the problem that many of these theories seem to be too linear in their flow of influence, rarely incorporating reciprocal feedback loops or external contingencies that might moderate the flow of influence. It is all a little bit too neat for our chaotic world. Third, few theories are specific about timelines, specifying how long it should take for a given process to affect some proximal indicator. Without such specifications, apparently disconfirming results make it difficult to know whether the next step in the model has not yet occurred or whether it will never occur because the theory is wrong in this particular. Fourth, the method places a great premium on knowing how to measure since failure to corroborate the model could result from partially invalid measures or from an invalid theory. Careful researchers can protect against this by developing more reliable measures. Fifth is the epistemological problem that many different models can usually be fit to any single pattern of data (Glymour, Sprites, Scheines 1987). The implication here is that causal modeling is more valid when multiple competing models are tested against each other rather than when a single model is tested.

The biggest problem though, is the absence of a valid counterfactual to inform us about what would have happened to students or teachers had there been no treatment. As a result, it is impossible to decide whether the observed data result from the intervention or would have occurred anyway. One way to guard against this is with "signed causes" (Scriven, 1976), predicting a multivariate pattern among the outcomes that is so unique it could only have occurred because of the reform. But signed causes depend on the availability of well-validated substantive theory and high quality measurement (Cook & Campbell, 1979). So, a better safeguard is to have at least one comparison group, and the best comparison group is a random one. So, we are back with random assignment and with the proposition that theory-based evaluations are useful complements to randomized experiments but not alternatives to them.

Conclusions

1. Random assignment is rare in research on the effectiveness of strategies to improve student performance, though it is widely acknowledged as the premier method for answering questions of this type. Educational evaluators do not do experiments. Substantive researchers in education do most of the experiments, though experiments are only a tiny fraction of all the cause-probing studies they do. The recent flurry of policy-relevant experiments has been done by scholars outside of Schools of Education.

2. Random assignment is common in pre-school education and also in schools when the concern is to prevent negative behaviors or feelings.

3. Intellectual culture may be one explanation for this disciplinary difference. Prevention researchers tend to be trained in public health and psychology where random assignment is esteemed and where funders and journal editors clearly prefer the

technique. Training and professional rewards are quite different for faculty in Schools of Education.

4. Capacity may be another explanation. Most school-based prevention experiments are shorter, implementation is by researchers rather than school staff, and the research topics may engage educators less than issues of school governance and teaching practice.

5. Nearly all educational evaluators believe that experiments are of little value. They believe that the theory of causation buttressing experiments is naïve, that experiments are difficult to implement, that they require unacceptable trade-offs, that they deliver a kind of information rarely used to change policy, and that the information they provide can be gained using simpler and more flexible methods.
6. Some of these beliefs are better justified than others. Beliefs about the viability of alternatives are particularly poorly warranted since no current alternative provides as convincing a causal counterfactual as the randomized assignment. However, the better criticisms suggest useful additions to school-based experiments, especially as concerns describing program theory, describing implementation quality, relating implementation to outcome changes, measuring whether theoretically specified intervening processes have occurred, and relating these intervening processes to student outcomes.
7. Random assignment is not a "gold standard". It creates only a probabilistic equivalence between groups, and then only at pretest. Moreover, treatment-correlated attrition is likely when treatments differ in intrinsic desirability. Also, treatments are not always independent, and the ways used to increase internal validity often reduce external validity.

8. More appropriate rationales for random assignment are that, even after all the limitations above are taken into account, (1) it still provides the logically most valid causal counterfactual; (2) it almost certainly provides a more efficient counterfactual; and (3) the results it generates are more credible in nearly all academic circles.
9. Educational evaluators will not be persuaded to do experiments simply by outlining their advantages and describing newer methods for implementing randomization. Most educational evaluators share some of the anti-experimental beliefs outlined above. To start a dialog, advocates of experimentation will need to be explicit about the method's limits and should seek to improve experimental practice by incorporating into it some of the critics' concerns about program theory, the quality of implementation, the value of qualitative data, the necessity of causal contingency, and meeting the information needs of school personnel as well as other stakeholders..
10. Even so, it will be difficult to enlist current educational evaluators behind a banner promoting more experimentation. However, they are not needed. They are not part of the controlled experimentation now beginning in schools, and future experiments can be carried out by substantive researchers in Education Schools and contract research firms, no to speak of university faculty in the policy sciences.

References

Angrist, J.D., Imbrens , G.W. & Rubin. D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, pp.444-462.

Anson, A., Cook, T.D., Habib, F., Grady, M.K., Haynes, N & Comer, J.P. (1991). The Comer School Development Program: A theoretical analysis. Journal of Urban Education, 26, pp 56-82.

Ares, C.E., Rankin, A. & Sturz, H. (1963). The Manhattan bail project: An interim report on the use of pre-trial parole. New York University Law Review, 38, 67-95.

Becker, H.S. (1958). Problems of inference and proof in participant observation. American Sociological Review, 23, pp. 652-660..

Berman, P. & McLaughlin, M.W. (1977). Federal programs supporting educational change. Vol. 8: Factors affecting implementation and continuation. Santa Monica, CA: Rand Corp.

Bhaskar, R. (1975). A realist theory of science. Leeds, England: Leeds.

Bogatz, G.A. & Ball, S. (1972). The impact of “Sesame Street” on children’s first school experience. Princeton, NJ: Educational Testing Service.

Boruch, R.F. (1974). Bibliography: Illustrated randomized field experiments for program planning and evaluation. Evaluation, 2, pp. 83-87.

Boruch, R.F. (1997). Randomized experiments for planning and evaluation: A practical guide. (Applied Social Research Methods Series, Volume 44.) Thousand Oaks, CA: Sage Publications.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

Campbell & Stanley (1963). Experimental and quasi-experimental designs for research. Chicago: Rand-McNally.

Campbell, F.A. & Ramey, C.T. (1995). Cognitive and school outcomes for high-risk African American students at middle adolescence: Positive effects of early intervention. American Educational Research Journal Vol 32, No. 4, pp 743-772.

Chelimsky, E. (1987). The politics of program evaluation. In D.S. Cordray, H.S. Bloom & R.J. Light (Eds.) Evaluation Practice in Review. San Francisco, CA: Jossey-Bass.

Chen, H. (1990). Theory-Driven Evaluations. Newbury Park, Ca: Sage, 10, pp. 95-103

Cohen, D.K. & Garet, M.S. (1975). Reforming educational policy with applied social research. Harvard Educational Review, 45.

Cicirelli, V.G. and Associates (1969). The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development, Vol. 1 and 2. A Report to the Office of Economic Opportunity. Athens: Ohio University and Westinghouse Learning Corporation.

Collingwood, R.G. (1940). An essay on metaphysics. Oxford: Clarendon.

Comer, J.P. (1988). Educating poor minority children. Scientific American, 259(5), pp. 42-48.

Connell, D.B., Turner, R.R. & Mason. (1985). Summary of findings of the school health education evaluation: Health promotion effectiveness, implementation and costs. Journal of School Health, 55, pp. 316-321.

Connell, J.P., Kubisch, A.C., Schorr, L.B. & Weiss, C.H. (Eds.). (1995). New approaches to evaluating community initiatives: Concepts, methods and contexts. Washington, D.C.: Aspen Institute.

Cook, T.D. (1985). Post-positivist critical multiplism. In R.L. Shotland & M.M. Mark (Eds.) Social science and social policy, pp. 21-62. Beverly Hills, CA: Sage.

Cook, T.D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. Sechrest and A.G. Scott (Eds.) New Directions for Program Evaluation: Understanding Causes and Generalizing about Them, Vol. 57. San Francisco: Jossey-Bass Publishers.

Cook, T.D. (2000). The False Choice between Theory-Based Evaluation and Experimentation. New Directions in Evaluation: Challenges and Opportunities in Program Theory Evaluation. 87, 27-34

Cook, T.D., Anson, A. & Walchli, S. (1993). From causal description to causal explanation: Improving three already good evaluations of adolescent health programs. In S.G. Millstein, A.C. Petersen & E.O. Nightingale (eds.) Promoting the Health of Adolescents: New Directions for the Twenty-First Century. New York: Oxford University Press.

Cook, T.D. & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Cook, T.D., Habib, F., Phillips, J., Settersten, R.A., Shagle, S.C., Degirmencioglu, S.M.(1999). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. American Educational Research Journal. 36(3), 543-597.

Cook, T.D., Hunt, H.D. & Murphy, R.F. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. American Educational Research Journal. 37(2), 535-597.

Corrin , W.J. & Cook, T.D. (1998). Design elements of quasi-experimentation. Advances in Educational Productivity, 7, pp. 35-57.

Cronbach, L.J. (1982). Designing evaluations of educational and social programs. San Francisco: Jossey-Bass.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F. & Weiner, S.S. (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Cronbach, L.J. & Snow, R.E. (1976). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington

Durlak, J.A. & Wells, A.M. (1997) (a). Primary prevention mental health programs for children and adolescents: A meta-analytic review. Special Issue: Meta-analysis of primary prevention programs. American Journal of Community Psychology, 25(2), pp. 115-152.

Durlak, J.A. & Wells, A.M. (1997) (b). Primary prevention mental health programs: The future is exciting. American Journal of Community Psychology, Vol. 25, No. pp. 233-241.

Durlak, J.A. & Wells, A.M. (1998). Evaluation of indicated preventive intervention (Secondary Prevention) Mental health programs for children and adolescents. American Journal of Community Psychology, Vol. 26, No. 5, pp 775-802.

Ehri, L., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z.. & Shanahan, T. (2001) Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. Reading Research Quarterly, 36, 250-287.

Ehri, L., Nunes, S., Stahl, S. & Willows, D. (2001) Systematic phonics instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. Review of Educational Research, 3, 393-447.

Elmore, R.F. (1996). Getting to scale with good educational practice. Harvard Educational Review, 66, 1-26.

Elmore, R.F. & McLaughlin, M.W. (1983). The federal role in education: Learning from experience. Education and Urban Society, 15, pp. 309-333.

Ferguson, R.F. (1998). Can Schools Narrow the Black-White Test Score Gap? In C. Jencks & M. Phillips. The Black-White Test Score Gap. Washington, DC: Brookings. pp. 318-374.

Fetterman, D.M. (Ed.), (1984). Ethnography in educational evaluation. Beverly Hills, CA: Sage.

Finn, J.D. & Achilles, C.M. (1990). Answers and questions about class size: A statewide experiment. American Educational Research Journal, 27(3), pp. 557-577.

Flay, B.R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Preventive Medicine, 15, pp. 451-474.

Fraker, T., & Maynard, R. (1987). Evaluating comparison group designs with employment-related programs. Journal of Human Resources, 22, 194-227.

Gasking, D. (1955). Causation and recipes. Mind, 64, pp. 479-487.

Gilbert, J.P., McPeek, B., & Mosteller, F. (1977). Statistics and ethics in surgery and anesthesia. Science, 198, 684-689.

Glass, G.V. & Smith, M.L. (1979). Meta-analysis of research on the relationship of class size and achievement. Educational Evaluation and Policy Analysis, 1, pp. 2-16.

Glymour, C., Sprites & Scheines, R. (1987). Discovering causal structure: Artificial intelligence, philosophy of science and statistical modeling. Orlando: Academic Press.

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

Greene, J.P., Peterson, P.E., Du, J, Boeger, L. & Frazier, C.L. (1996). The effectiveness of school choice in Milwaukee: A secondary analysis of data from the program's evaluation. University of Houston mimeo. August, 1996.

Guba, E.G. & Lincoln, Y. (1982). Effective evaluation. San Francisco: Jossey-Bass.

Gueron, J. M. (2002). The politics of random assignment: Implementing studies and affecting policy. In F. Mosteller & R. Boruch (Eds). Evidence Matters: Randomized Trials in Education Research. Washington, D.C.: Brookings Institution Press.

Hanushek, E.A. (1999). Evidence on class size. In S. Mayer & P. E. Peterson (Eds.) When schools make a difference. Washington, D.C.: Brookings Institute.

Herrman, R. (1998). Whole school reform: A review. American Institutes for Research. Washington, D.C.

Holland, P.W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, pp. 945-970.

House, E. (1993). Professional evaluation: Social impact and political consequences. Newbury Park, CA:., Sage Publications.

Howell, W.G., & Peterson, P.E. (2002) The Education Gap: Vouchers and Urban Schools. Washington, D.C.: Brookings Insitution Press.

Kemple, J.J. (2001) Career academies: Impacts on students' initial transitions to post-secondary education and employment. New York: Manpower Demonstration Research Corporation.

Kuhn, T.S. (1970). The structure of scientific revolutions. (2nd edition) Chicago: University of Chicago Press.

LaLonde, R.J. (1986). Evaluating the econometric evaluations of training programs with experimental data. American Economic Review, 76, 604-620.

Latour, B. & Woolgar, S. (1979). Laboratory life: The construction of scientific facts. Beverly Hills: Sage Publications

Leviton, L.C. & Cook, T.D. (1983). Evaluation findings in education and social work textbooks. Evaluation Review, 7, pp. 497-518.

Lindblom, C.E. & Cohen, D.K. (1980). Usable knowledge. New Haven, CT: Yale University Press.

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

Mackie, J.L. (1974). The cement of the universe. Oxford, England: Oxford University Press.

Mosteller, F., Light, R.J. & Sachs, J.A. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. Harvard Educational Review, 66, pp. 797-842.

Nave, B., Miech, ElJ. & Mosteller, F. (1999). A rare design: The role of field trials in evaluating school practices. Paper presented at The American Academy of Arts and Sciences an Harvard University.

Olds, D.L., Eckenrode, D., Henderson, C.R., Kitzman, H, Powers, J. Cole, R., Sidora, K., Morris, P., Pettitt, L.M.& Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. Journal of the American Medical Association, 278(8): 637-643.

Peters, R.D.& McMahon, R.J. (Eds.) (1996). Preventing childhood disorders, substance abuse, and delinquency. Banff International Science Series, Vol. 3, pp 215-240. Thousand Oaks, CA: Sage.

Peterson, P.E., Myers, D. & Howell, W.G. (1998). An evaluation of the New York City School Choice Scholarships Program: The First Year. Paper prepared under the auspices of the Program on Education Policy and Governance, Harvard University. PEPG98-12

Peterson, P.E., Greene, J.P., Howell, W.G. & McCready, W. (1998). Initial findings from an evaluation of school choice programs in Washington, D.C. Paper prepared under the auspices of the Program on Education Policy and Governance, Harvard University.

Popper, K.R. (1959). The logic of scientific enquiry. New York: Basic Books.

Quine, W.V. (1951). Two dogmas of empiricism. Philosophical Review, 60, pp. 20-43.

Quine, W.V. (1969). Ontological relativity and other essays. New York: Columbia University Press.

Raikes, H.H. & Love, John. M. (2002) Early Head Start: A dynamic new program for infants and toddlers and their families. Infant Mental Health Journal, 23, 1-13.

Ramey, C.T. & Campbell, F.A. (1991). Poverty, early childhood education, and academic competence: The Abecedarian experiment. In A.C. Huston (Ed.) Children in poverty: Child development and public policy. New York: Cambridge University Press, pp. 190-221.

Riecken, H.W. & Boruch, R. (1974). Social experimentation. New York: Academic Press.

Rivlin, A.M. & Timpane, M.M. (Eds.) (1975). Planned variation in education. Washington, DC: Brookings Institution.

Rossi, P. H. & Lyall, K. C. (1976) Reforming public welfare. New York, N.Y. Russell Sage Foundation .

Rouse, C.E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee parental choice program. The Quarterly Journal of Economics, pp. 553-602.

St.Pierre, R.G., Cook, T.D. & Straw, R.B. (1981). An evaluation of the Nutrition Education and Training Program: Findings from Nebraska. Evaluation and Program Planning, 4, pp. 335-344.

Schweinhart, L.J., Barnes, H.V. & Weikart, D.P. (with W.S. Barnett and A.S. Epstein) (1993). Significant benefits: The High/Scope Perry Preschool Study through age 27. Ypsilanti, MI: High/Scope Press.

Scriven, M. (1976). Maximizing the power of causal investigation: The Modus Operandi method. In G.V. Glass (Ed.), Evaluation Studies Review Annual, 1, pp. 101-118. Newbury Park, CA: Sage Publications.

Shadish, W.R. & Cook, T.D. (1999) Design rules. Statistical Science, 294-300.

Shadish, W.R., Cook, T.D., & Campbell D.T.(2001). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghtin-Mifflin.

Stake, Robert E. (1967). The countenance of educational evaluation. Teachers College Record, 68, 523-540.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Anderson, R.B. & Cerba, T.R. (1978). An evaluation of Follow Through. In T.D. Cook (Ed.) Evaluation Studies Review Annual, 3, pp. 571-610. Beverly Hills CA: Sage Publications.

Vinovskis, M.A. (1998). Changing federal strategies for supporting educational research, development and statistics. Background paper prepared for the National Educational Research Policy and Priorities Board, U.S. Department of Education.

Vinoskis, M.A. (2002). Missing in practice: Development and evaluation at the U.S. Department of Education In F. Mosteller & R. Boruch (Eds). Evidence Matters: Randomized Trials in Education Research. Washington, D.C.: Brookings Institution Press.

Wargo, M.J., Tallmadge, G.K., Michaels, D.D., Lipe, D. & Morris, S.J. (1972). ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970, Final Report, Contract No. OEC-0-71-4766. Palo Alto, CA: American Institutes for Research.

Weiss, C.H. (1988). Evaluation for decisions: Is anybody there? Does anybody care? Evaluation Practice, 9, pp. 5-20.

Weiss, C.H. & Bucuvalas, M.J. (1977). The challenge of social research to decision making. In C.H. Weiss (Ed.), Using social research in public policy making. pp. 213-234. Lexington, MA: Lexington.

Whitbeck, C. (1977). Causation in medicine: The disease entity model. Philosophy of Science, 44, pp. 619-637.

Witte, J.F. (1998). The Milwaukee voucher experiment. Educational Evaluation and Policy Analysis, 20, pp. 229-251.

Zdep, S.M. (1971). Educating disadvantaged urban children in suburban schools: An evaluation. Journal of Applied Social Psychology, pp. 1,2, 173-186