"A Critical Appraisal of the Case Against Using Experiments to Assess School (or Community) Effects"

Tom Cook

Professor of Sociology
Northwestern University
Introduction

Compared to most industrialized nations, the American system of primary and secondary education is uniquely decentralized. Federal, state and local authorities have a say in educational policy, with the state and local roles being most influential. As a result, educational goals and practices vary enormously, not only by state and school district, but also by schools within districts. To many foreigners, the American system must look like a cacophony of local experimentation. And it has probably come to look even more so in the last thirty years. Calls have emanated from politicians, business leaders and educational policy pundits seeking to improve schools through identifying effective school practices, setting higher standards, increasing accountability, developing more school-based management, founding charter schools, distributing school vouchers, creating smaller schools and class sizes, using more and newer technologies, and instituting more and better teacher and principal training. All this local experimentation indicates to some commentators a vibrantly democratic school system, free of the centralized control found elsewhere (e.g., Louis, 1998).

But experimentation connotes more than implementing different ways of doing things. It also connotes systematically evaluating these alternatives--usually through deciding which alternatives are to be compared, which criteria they are to be compared on, how data on each criterion are to be collected, and how a decision about relative effectiveness is to be reached once all the alternatives have been compared on all the criteria. To scholars, experimentation further connotes: (1) studies that take place in laboratories from which all theoretically irrelevant causal forces have been excluded and in which the experimenter has well-nigh total control over when and how a causal agent is manipulated; or (2) using random assignment for deciding which units--usually schools, classrooms or students in educational work--are to be exposed to the various treatment alternatives under test. This is to rule out the possibility that any observed post-assignment group differences are due to pre-existing group differences rather than to differences in the treatment experienced.
Education uses experiments in both these senses. However, our focus on evaluating reforms in school settings inclines us to be concerned only with random assignment. Yet such assignment is rare in the most successful laboratory sciences, indicating that it is not logically necessary for strong causal inferences. However, it is the best single facilitant of such inferences in the many open system contexts where the cause or effect are not well understood theoretically, as with school reform. Commentaries on the superiority of random assignment for drawing causal inferences in non-laboratory settings are routine in the philosophy of science and in method texts in health, public health, agriculture, statistics, micro-economics, psychology and those parts of political science and sociology dealing with improving the assessment of public opinion. The same advocacy is evident in elementary method tests in education, though not—as we shall see—in specialized texts on evaluating educational reforms.

If the American education system promotes experimentation in the sense of implementing many variants, it does not promote experimentation in the sense of using random assignment to assess how effective these variants are. Nave, Meich and Mosteller (1999) showed that not even 1% of dissertations in education or of the studies archived in ERIC Abstracts involved randomized experiments. Casual review of back numbers of the premier journals in the field tell a similar story, whether for the American Educational Research Journal or Educational Evaluation and Policy Analysis. Since my interest is in whole school reform rather than changing discrete features within schools—like curricula—I wrote to a colleague with a national reputation who designs and evaluates curricula. She replied that in her area randomized experiments are extremely rare, adding "You can't get districts to randomize or partially adopt after a short pilot phase because all parents would be outraged (from either side)."

As further evidence of the rarity of controlled experiments, consider specific areas of whole school reform of interest today. A review of school desegregation studies organized by the then National Institute of Education (1984) showed just one such experiment (Zdep, 1971). I know of no experiments on standards setting. The effective schools literature reveals no experiments where the
school practices presumed to be effective from correlational studies were then implemented in some schools at random and withheld from others. The plethora of 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, 1999). This suggests there are no such experiments on the effects of Catholic or Accelerated or Total Quality Management schools. On vouchers I know of only one completed study, often reanalyzed (Witte, 1998; Greene, Peterson, Du, Boeger & Frazier, 1996; Rouse, 1998) and of two others now underway (Peterson, Greene, Howell, & McCready, 1998; Peterson, Myers & Howell, 1998). On charter schools I know of no relevant experiments. On smaller class sizes (where classes rather than schools are the unit of assignment), I know of six experiments, the most recent and 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, currently underway (Kemple & Rock, 1996). While on teacher training I know of no relevant studies where the school is the unit of assignment. So, current knowledge of effective educational policy concerning elementary and secondary schools has to depend on methods other than random assignment.

Equally as striking is that, of the few experiments cited above, nearly all were conducted by scholars whose primary organizational affiliation is outside of education. The best-known class size experiment was begun by educators (Finn & Achilles, 1990), but popularized by statisticians (Mosteller, Light & Sachs, 1996). The Milwaukee voucher study was done by political scientists (Witte, 1998) and reanalyzed by political scientists (Greene, et al 1996) and economists (Rouse, 1998). The Comer studies were conducted by sociologists and psychologists (Cook et al, 1999; Cook, Hunt & Murphy, 1999). The ongoing experiment on academies within high schools (Kemple & Rock, 1996) is being done by economists, and the work on school choice programs in Washington, DC and New York is also being done by political scientists (Peterson et al, 1998). Striking is the twenty-year paucity of experimental studies conducted by scholars with appointments in Schools of Education. Yet Schools of Education are precisely where we might expect the strongest evaluations of school reform to be done.
That they are not done there does not reflect ignorance. The vast majority of writers on educational research methods understand the logic of experimentation, know of its theoretical benefits, and appreciate how esteemed it is in science writ large. However, most (including Alkin, Cronbach, Eisner, Fetterman, Fullan, Guba, House, Hubermann, Lincoln, Miles, Provus, Rist, Sanders, Schwandt, Stake, Stufflebeam and Worthen) denigrate formal experiments at some time in their writings. A few pay lip-service to experiments while carrying out non-experimental studies of their own whose virtues they therefore model for others (e.g. Bryk, Porter, Scriven and C.H. Weiss).

Such distaste for experiments stands in stark contrast to what we find among scholars who do empirical work in schools but without operating out of a School of Education. Foremost among them are those scholars who seek to learn about ways to improve student mental health and to prevent violence or the use of tobacco, drugs and alcohol (e.g., Durlak & Wells, 1997; Peters & McMahon, 1996; Cook, Anson & Walchli, 1993). They routinely assign schools to treatments at random. In a personal conversation Brian Flay--Director of the Prevention Research Center at the University of Illinois, Chicago--estimated the number of such experiments to be in the hundreds. More careful analysis of the Durlak & Wells meta-analysis should enable a more accurate estimate. It is also worth noting that a few researchers investigating curricula designed to improve children's knowledge of nutrition and cafeteria eating choices have also conducted randomized experiments (St. Pierre, Cook & Straw, 1981; Connell, Turner & Mason, 1985) True experiments are commonplace in some areas of contemporary research on primary and secondary schools. But they are not being done by researchers exposed to methodology or evaluation training in Schools of Education. As a result, few controlled experiments are available on the topics of greatest interest to school policy debates (e.g., on school governance, school structure, school funding patterns as with vouchers, the academic curriculum and professional training).
The present paper seeks to demonstrate what is special about the intellectual culture of research on school reform in those Schools of Education where research is routinely conducted—a culture that actively rejects random assignment in favor of alternatives that the larger research community judges to be technically inferior. Educational researchers have lived for at least 20 years knowing they have rejected what science apotheosizes. They will not adopt experimentation merely by learning more about its benefits, about the shortfalls of the alternatives they prefer, or about why experiments are more feasible today than 20 years or so ago. They believe in a set of mutually reinforcing propositions that provide them with what they believe is an overwhelming rationale for rejecting experiments on any number of ontological, epistemological, methodological, practical or ethical grounds. The writers we cite differ in some of the reasons they offer for rejecting a scientific model of knowledge growth. But any Ph.D. from a School of Education exposed to the relevant literature on evaluation methods would have encountered arguments against experiments that appeared cogent and comprehensive. And for more mature researchers, the recent call to conduct formal experiments may have a “deja vu” quality, reminding them of a battle they thought they had won long ago—the battle against a “positivist” view of science that privileges the randomized experiment and the research and development model related to it with origins in agriculture, health, public health, marketing or even the military. They think this model is irrelevant to the special organizational complexity of schools, and they prefer an R&D model based on management consulting.

In management consulting, the crucial assumptions are (1) that each organization (e.g. school or school district) is so complex in its structure and functions that it is difficult to implement any changes that involve the organization’s central functions; (2) that each organization is unique in its goals and culture, entailing that the same stimulus is likely to elicit quite variable responses depending on a school’s history, organization, personnel and politics; and (3) that suggestions for change should be reflect the creative blending of knowledge from many different sources—from general organizational theories, from deep insight into the district or schools under case study, and from “craft” knowledge of what is likely to improve this particular school or district (Lindblom & Cohen, 1980). Scientific knowledge about
effectiveness is not particularly prized in this account, especially if it is produced in settings different from those where the knowledge is to be applied. So, random assignment is seen as a central component of an inappropriate scientific world view that obscures each school’s uniqueness, that oversimplifies the multivariate and non-linear nature of full explanatory causation, and that is naive about the mostly indirect ways in which social science is used in policy debates. Most educational evaluators see themselves at the forefront of a post-positivist, democratic and craft-based model of knowledge growth that is superior to the elitist scientific model they think has failed to create useful and valid knowledge about improving schools.

In what follows we describe the major beliefs education researchers hold that undermine faith in random assignment. We also discuss the validity and relevance of these beliefs, finding them to vary considerably in credibility. We take some of the objections very seriously, however, and conclude they lead to a subtle modification of the usual rationale for random assignment; but without vitiating its utility. The valid objections speak to important research tasks for which random assignment is irrelevant and perhaps even sometimes harmful. So, we outline appropriate roles within experiments for the kinds of inquiry that many education researchers now prefer as alternatives to random assignment. We suggest that unless the concerns of critics of random assignment are incorporated into a broader framework of research design, we cannot hope to see more experiments being done either by those who currently teach methods for evaluating educational reforms or by their students. Nor can we expect much support for random assignment among those mid-level employees in government, foundations and contract research firms who have been influenced by the current crop of educational evaluators. While the support of such groups is not necessary for increasing the number of controlled experiments, we argue that taking their concerns into serious account will increase the yield from experiments, whatever the disciplinary affiliation of those who do them.

Beliefs Adduced to Reject Random Assignment
The causal world is ordered more complexly than a causal connection from A to B can represent. With a few exceptions (e.g., Guba & Lincoln, 1982), educational researchers seem willing to postulate that a real world exists, however imperfectly it can be known. For any given outcome, randomized experiments test the influence of only a small subset of potential causes, often only one. And at their most elegant they can responsibly test only a modest number of interactions between different treatments or between any one treatment and individual differences at the school, classroom or individual level. Thus, randomized experiments are best when a causal question is simple, sharply focused and easily justified.

The theory of causation most relevant to this conception of cause has been variously described as the manipulability, activity or recipe theory (Collingwood, 1940; Gasking, 1955; Whitbeck, 1977). It places primacy on identifying the consequences of discrete activities that can be brought under human will and actively manipulated, hopefully in multiple settings so as to generate a causal recipe that will dependably bring about a desired change. Although the goal is to describe the causal consequences of a given treatment, there are conditions under which these descriptions can aid explanation (Mackie, 1974). This is especially when the manipulations help discriminate between competing theories--an easier proposition in fields with rich substantive theory. However, by itself random assignment is irrelevant to explanation; it only helps describe the effects of some deliberately varied event.

Contrast this with the theory of causation most esteemed in science. In some form or another, the preference is for full explanation in either of two related senses. One emphasizes “generative processes” (Bhaskar, 1975) that can bring about effects in a wide variety of circumstances--like gravity as if affects falling, or a specific genetic defect as it induces phenylketonuria, or time on task as it facilitates learning. But even here the relationships are contingent--gravity works only so long as there is not a vacuum; the genetic defect does not induce phenyletonuria given certain childhood diets; time on task does not induce learning if a student is not engaged or if the curriculum materials are meaningless.
So, a second and more general understanding of causation is that all the contingencies are specified that impact on a given consequence or that follow from a particular event (Mackie, 1974).

Cronbach (1980) postulates that in the real world of education multiple causal factors are implicated in any desired change in students or teachers, further contending that these change factors are often not linearly related to each other. His model of real world causation is more akin to a pretzel (or intersecting pretzels) than to any simple arrow from A to B. He cannot imagine an educational intervention that fully explains an outcome--at most it will be just one determinant--or an intervention that is so general in its effects that the size of a cause-effect relationship remains constant across different populations of students and teachers, across different kinds of schools, across the other novelties simultaneously occurring in schools, across the entire range of relevant outcomes, and across all historical time periods. Causal contingency is the watchword for anyone interested in full explanatory causation, making the paucity of contingencies implicit in the activity theory of causation a serious limitation--a paucity that is the product of the small number of variables that can be manipulated, and of the small number of planned interactions that can be examined in one experiment. Experiments cannot faithfully represent a real world characterized by multivariate, non-linear (and often reciprocal) causal relationships so that, seated in an armchair or around the table in some foundation, few educational researchers have much difficulty detailing contingencies likely to limit the effectiveness of any proposed intervention. Indeed, Cronbach and Snow (1976) gained considerable visibility by initiating the search to discover, not treatments with robust main effects, but aptitude by treatment interactions where a treatment’s effect varies with student or teacher characteristics. There is definitely substance to the notion that randomized experiments speak to a simple, and possibly oversimplified, theory of causation.

But Cronbach and Snow were not able to find many robust interactions in the educational literature. Moreover, many educational researchers speak and write as though they accepted certain uncontingent causal connections--e.g., that small schools are better than large ones; that time-on-task raises achievement; that summer school raises test scores; that school desegregation hardly affects
achievement; and that assigning and grading homework raises achievement. And they also seem to be willing to accept some propositions with very little causal contingency—e.g., reducing class size increases achievement provided that it is a “sizable” change and to a level under 20, and that Catholic schools are superior to public ones in inner-city but not suburban settings. Commitment to a full explanatory theory of causation has not precluded some educational researchers from acting as though the United States educational system can be characterized in terms of some dependable main effects and some very simple non-linearities whose form is very much like arrows flying from A to B (or at least from A and C to B).

Moreover, experiments were not designed to meet grand explanatory goals. Nor were they originally designed for causal verisimilitude. Rather, they were primarily designed to pull apart what is often confounded in Nature so as to better examine a causal link. Yet the move to take experimental methods outside of the laboratory and into schools (or fields, hospitals and doctors’ offices) implies that verisimilitude is not irrelevant in the Pantheon of experimental desiderata. However, it is always secondary to achieving as clear a picture as possible of the extent to which the link between a presumed cause and effect is causal in the activity theory sense (Campbell, 1957; Campbell & Stanley, 1968). So, a special onus falls on advocates of experimentation. They have to make the case that any treatment is so important to policy or theory, so likely to influence the outcome at the margin, and so likely to be generalizable in its effects that it is worthwhile to invest in an oversimplified theory of causation and to accept any restrictions to verisimilitude that might follow from assigning units at random—of which more later.

Some causal contingencies are not very relevant to educational policy. In school research special emphasis deserves to be placed on identifying those causal contingencies that modify the sign of a causal relationship and not just its magnitude. Identifying causal sign switches informs us where a treatment might be directly harmful as opposed to simply having less positive benefits for some children than others. This is an important distinction, for it is not always possible to assign different treatments to
different populations of schools, students or teachers. Sometimes, policy-makers are more than willing to advocate a change that works differentially across student groups, so long as the effects are rarely negative. So, not all factors moderating a cause-effect relationship are equally important in a practical sense. While important to full explanation, third variables that slightly modify the value of slopes but not their sign may be of lesser use in educational policy.

It is important not to forget how modest experimentation is in its goals, however important they are. The theory of causation undergirding random assignment is very simple when compared to grand explanatory theories; experimenters are forced to attribute primacy to those causal relationships that seem to be good contenders as generalized main effects; they have to overlook many of the possible contingency variables that affect the size of a causal relationship, but not its sign; and they have to acknowledge that any causal dependability they detect will be probabilistic rather than deterministic, since all causal connections are embedded within more complex explanatory systems where different kinds of schools, students, settings, etc. can modify effect sizes. In education, experiments are done to learn about the consequences of quite practical multivariate treatment packages, and student impact takes precedence over theory-testing. Who can entirely blame opponents of experimentation who believe that it is better to have a more biased answer to a big explanatory question than to have a less biased answer to a smaller descriptive question? We disagree with the judgment that learning about the consequences of educational reforms involves small questions. But it is clear that scholars of all kinds take explanation to be their Holy Grail; not identifying the consequences of some multivariate treatment package designed more for practical impact than theory development. We proponents of random assignment should not over claim for our preference.

Random assignment depends on “positivist” epistemological premises that have been discredited. To philosophers of science positivism connotes a rejection of realism, the formulation of theories in mathematical form, the requirement that theories predict some phenomenon perfectly without necessarily "explaining" it, and a belief in definition operationalism--that is, IQ has no independent entitity; it is only
what an IQ test measures. This epistemology has been discredited since the 1940’s, but many educational researchers still invoke “positivist” in a less specific sense that includes all quantitative research or all hypothesis-testing or both. Since randomized experiments nearly always involve hypothesis tests and quantitative data manipulation, to reject such procedures entails rejecting experiments.

Kuhn’s work (1970) stands at the forefront of the reasons cited for rejecting “positivist” science. He argued two things of relevance. First, that theories cannot be formulated so specifically that definitive falsification results. This is his claim about the “incommensurability of theories”. And second, he argued that all measures are impregnated with the theories, hopes, wishes and expectations of investigators, undermining their neutrality for discriminating between truth claims. This is his claim about the “theory-ladenness of observations”. In destroying the idea of totally explicit theories and totally neutral observations, Kuhn seems to have undermined the rationale for science in general and for random assignment in particular. After all, the latter uses observational data to test causal hypotheses. The work of writers on educational evaluation like Cronbach, Eisner, Fetterman, Guba, House, Lincoln, Stake and Stufflebeam are filled with references to Kuhn and to philosophers with somewhat similar views like Lakatos, Harré and Feyerabend. Also mentioned are those iconoclastic sociologists of science whose studies of laboratory behavior reveal practicing scientists whose on-the-job behavior deviates markedly from scientific norms. All this is designed to show that modern meta-science has undressed the emperor of science. Indeed, some educational researchers probably believe that recent advocacies of random assignment emanate from naked emperors who think they are wearing the most beautiful clothes that science has yet created but who are, in reality, wearing almost transparent and very much recycled old experimental tatters.

This epistemological critique is overly simplistic. Even if it is true that theories can never be totally explicit and observations never theory-neutral, this does not negate the idea of progress in science. Many observations have stubbornly reoccurred across the very many perspectives researchers
have brought to bear on a problem. Indeed, as theories replace each other, most of the fact-like statements from the older theory are incorporated into the newer one, stubbornly surviving whatever the theoretical superstructure. It is probably true that there are no “facts” we can independently know to be certain; but there are many propositions with such a high degree of facticity that they can be confidently treated as though they were facts.

For practicing scientists, including experimenters, the trick is to make sure that observations are not impregnated with a single theory. This means assessing one’s observations from many different theoretical positions, including those of one’s theoretical opponents. It also leads to assigning a high value to independent replications of all or part of a causal claim, especially if theoretical opponents do them. Moreover, there are many ways in which experimenters can increase the facticity of their own work, primarily by making more heterogeneous the points of view built into the research and by ensuring to the extent possible that these sources of heterogeneity do not all share the same direction of bias (Cook, 1985). Kuhn complicates what a “fact” means; but he does not deny that some claims to a fact-like status are stronger than others, particularly those based on relationships and readings that stubbornly reoccur whatever the predilections of researchers.

It is also likely that theoretical statements are never definitively tested (Quine, 1951; 1969), including mundane statements about the effects of some educational program. But this does not mean that individual experiments fail to probe theories and causal hypotheses. Thus, when the results of a study are negative, program developers (and others) are likely to surface methodological and substantive contingencies that could have brought about a different result--perhaps with a different measure or a different group of students. Subsequent studies then probe these contingency formulations and, if they again prove negative, lead to the next round of probes of whatever more complicated contingency hypotheses program developers may have come up with to explain the second round of disconfirmations. After a time, the process understandably runs out of steam, so particularistic are the contingencies that remain to be examined. It is as though most of the scholarly community concludes:
“Yes, the program might be effective under some rare and as yet unexamined contingencies. But it has
not been effective under many other conditions, and those that remain to be examined are so
circumscribed that the reform option cannot be worth much even if it is effective under the as yet
unexamined conditions.” The advocates of Kuhn are correct, though. The process I am describing is
part social, not exclusively logical; and the dilemma arises because the underlying program theory is not
sufficiently explicit that it can be definitively confirmed or rejected in a single study or even a program of
studies. But this elasticity of theory does not mean that decisions about the viability of a causal
hypothesis are only social and that they are devoid of all empirical and logical content.

This response to the positivist critique stresses the need for program of research, including
experimental research. A one-shot study is not likely to set to rest all concerns about possible causal
contingency, whatever the claims originally made for single studies in the purest falsificationist
approaches to theory-testing (Popper, 1959). Taking Kuhn seriously implies the need for a program of
experimental research on almost any topic. However, critics have not been interested in developing a
better, non-positivist epistemological justification for experimentation, possibly seeing such efforts as
futile band-aids patching up a doomed theory of experimental verification and falsification. Instead, they
have turned their attention to developing post-positivist theories of methods for learning about
educational reforms that (1) stress qualitative methods and hypothesis discovery over quantitative
methods and hypothesis testing (e.g. Guba & Lincoln, 1982; Cronbach, et al 1980; Cronbach, 1982);
or (2) measure the extent to which the mediating processes specified in the substantive theory of a
program have actually occurred in the time sequence postulated (Connell, Kubisch, Schorr & Weiss,
1995; Chen & Rossi, 1987). In neither case is there an emphasis on assessing performance relative to
a valid causal counterfactual––the crucial function of random assignment. But before turning to these
alternative methods, we have to acknowledge the more specific attacks made on random assignment
per se.
Random assignment has been tried and has failed on its own terms. Education researchers were at the forefront of the flurry of social experimentation that took place at the end of the 1960’s and the 1970’s. Studies of the effects of Head Start (Cicirelli & Associates, 1969), Follow Through (Stebbins, St. Pierre, Proper, Anderson & Cerva, 1978), Title I (Wargo, Tallmadge, Michaels, Lipe & Morris, 1972) and Sesame Street (Ball & Bogatz, 1970; Bogatz & Ball, 1971) became available, the first three concluding that there were no effects of any magnitude or replicability. The results were greeted with considerable dispute about the methods and forms of analysis used, and many educational evaluators concluded from this that quantitative evaluation of all kinds had failed. So, they turned to other methods, propelled in the same direction by their reading of Kuhn. Other scholars responded differently, coming to emphasize the study of school management and program implementation in the belief that poor management and implementation were part of the reason for the disappointing results achieved (Berman & McLaughlin, 1977; Elmore & McLaughlin, 1983; Cohen & Garet, 1975). In any event, dissatisfaction with quantitative evaluation methods grew.

However, none of the most heavily criticized studies had involved random assignment. In education, criticism of the capacity of true experiments to deliver what they promised was a task that Cronbach and his coauthors (1980) took on. They reanalyzed some of the experiments from the lists Riecken & Boruch (1974) and Boruch (1974) had generated in order to counter arguments about the infeasibility of conducting randomized experiments in extra-laboratory settings. Cronbach paid particular attention to the Vera Institute’s Bail Bond Experiment and the Negative Income Tax Experiments, and was able to show to his own and his followers’ satisfaction that these studies were flawed in how they were implemented as randomized experiments. Hence, for this (and many other reasons), he believed they did not warrant the conclusions the original investigators had reached and could not constitute a valid model for evaluating educational innovations. So, the belief grew in education that many studies presented as randomized experiments were in fact flawed for learning about meaningful causal relationships.
However, so few randomized experiments were available in education at the time that the studies Cronbach analyzed were from other fields. I know of only three randomized experiments on educational reform available at the time. One was of the second year of Sesame Street (Bogatz & Ball, 1971) where cable capacity was randomly assigned to homes in order to promote differences in children’s opportunity to view the show. (To increase the odds even further, children in the cable condition were also regularly visited by research staff who left behind toys, books and games about the show so that the operational treatment was both viewing and social encouragement to view.) A second experiment was the Perry Preschool Project (Schweinhart, Barnes & Weikart, 1993) and the third involved only 12 youngsters randomly assigned to be in a desegregated school (Zdep, 1971). Only Zdep's study involved primary or secondary schools, and so it was probably not accurate to claim in the 1970’s that randomized experiments had been tried in education and had failed to be properly implemented there. Only non-experimental quantitative studies had been done---on school desegregation, for example---and none of these would pass muster even as quality quasi-experiments (Cook, 1984).

But this brief retrospective should not be taken to imply that it is easy to implement randomized experiments of school reform. Serious implementation difficulties have arisen in many of the more recent experiments. First, cases have been documented of schools dropping out of treatment conditions in different proportions, largely because a new principal wants to change what his or her predecessor recently did, including the activities defining a treatment that comparison groups do not experience (e.g., Cook, Hunt & Murphy, 1999). Moreover, in the Tennessee class size study it is striking that the number of classrooms in the final analysis differs by more than 20% across the three treatment conditions, even though the randomized design used should have resulted in similar numbers. Then, there are the cases of inadvertent treatment crossovers, as happened in Cook et al (1999) in Prince George’s County. One principal in a treatment condition was married to someone teaching in a control school; one control principal really liked the treatment and learned more about it for himself and tried to implement parts of it in his school; and a teacher in one control school was the daughter of one of the central program
officials working at Yale and he several times came to her school to talk to the whole school about how to create a better school. The crossover involved only three of 23 schools and none of the three received the central treatment components--a school-based facilitator and training procedures at Yale and in the district. Still, the planned experimental contrast was diluted to some unknown degree. In a similar vein, the Tennessee class size experiment compared classrooms within schools, though most public health work on prevention uses between-school designs to minimize the chances of treatment contamination. What did Tennessee teachers in the larger classes make of the situation whereby some colleagues in the same school taught smaller classes at the same grade level? Were they dispirited and so tried less? Few randomized experiment in education can escape from issues like these about how random assignment and treatment independence were maintained. But, when implemented and monitored with care, random assignment is sometimes feasible in schools and can sometimes be maintained over time. The public health work on prevention attests to this.

It is not easy to explain why randomization is more successful on public health topics in schools. Some reasons are surely organizational, having to do with the priority public health researchers have learned to attribute to clear causal inferences--a priority reinforced by their funders (mostly NIH, CDC and the Robert Wood Johnson Foundation). The appropriate contrast is with the lower (almost non-existent) priority accorded to random assignment by federal and state Offices of Education and by the many foundations funding education reform, though we were very recently heartened by a small program announcement joint between NSF, OE and NICHD whose “long-term goal... is to develop the knowledge and experimental methods that will allow for the implementation and evaluation of large-scale educational interventions”. Other reasons for the discipline differences in the frequency of conducting experiments have to do with substantive matters. The public health work is mostly about curricula rather than, say, whole school reform; and the curricula are circumscribed in time, rarely lasting even a single school year. Moreover, unlike in math, science or reading, teachers do not have to be trained to deliver the experimental materials. Researchers typically do it. Hence, the implementation shortfall is less than when teachers implement and have not mastered all the complexities of a new way
of doing things. Finally, we should not forget that the prevention work requires no new forms of coordination among administrators and teachers, among different teachers, or between teachers and parents.

The very real difficulties of random assignment for studying school reform suggest a subtle but important modification to its usual rationale—a modification that should bolster the credibility of advocates of random assignment by making it clear they are not proposing a causal “gold standard”. Theoretically, the rationale for random assignment is beyond dispute. An expected pre-treatment difference is created between groups that is zero. But causal interpretation depends on a zero post-treatment difference between groups in all things other than treatment assignment, as well as on the absence of treatment crossovers. In the real world, selective attrition and treatment crossover sometimes occur, however sophisticated attempts might have been to reduce their occurrence. Is it not a more appropriate rationale to argue that, in actual research practice, random assignment creates a better counterfactual than any of its likely alternatives based on self- or administrator selection into treatments? This rationale does not require a perfect counterfactual, and enunciating it encourages all experimenters to check how well random assignment and treatment independence have been achieved and maintained. Certain knowledge of causal connections does not inevitably follow from random assignment, though better knowledge does. Appearing to argue for perfection raises a red flag that infuriates critics in education.

**Random assignment is neither politically, administratively nor ethically feasible in education.** The small number of randomized experiments in education may reflect, not researchers’ distaste for them, but a simple calculation of how difficult they are to mount in the complex organizational context of schools. School district officials do not like the focused inequities in schools structures or resources that random assignment sometimes generates, fearing negative reactions from parents and school staff. They prefer individual schools to choose which reforms they will implement or to make changes on a district-wide
basis. Principals and other school staff probably share these preferences and have additional administrative concerns about disrupting routines when trying something new. Inevitably, this disruption will be greater than in the control condition, and more so when whole school reforms are at issue. And finally, there are the usual kinds of ethical concerns about withholding potentially helpful treatments from students and teachers in need. Given such objections, it would need people committed to random assignment and armed with the widely available counter-objections (e.g., Boruch, 1998) to fight for it. But since most evaluators in education do not believe that random assignment is either feasible or valuable, they are not likely to fight for its expansion, given what they currently know.

What does it take to mount randomized experiments, even for whole school reforms? In the Cook et al (1999) study in Prince George’s County, Maryland, random assignment was sponsored by the school district and all district middle schools had to comply. So, principals had no choice over participating in the study or in the treatment they eventually received. The district took this step because a foundation-funded network of very prestigious scholars--none from education-- insisted on random assignment as a precondition for funding the program and its evaluation. In a second case evaluating the same program in Chicago (Cook, Hunt & Murphy, 1999), it was the principal investigator who insisted on random assignment as a precondition for collaborating with the program implementers. Moreover, in deliberate contrast to the Maryland experiment the Chicago study was restricted to schools where all the principals wanted the program but said they were prepared to live with the results of the coin toss and to tolerate the annual questionnaire measurement of students and staff, irrespective of the treatment to which they would eventually be assigned. Random assignment of schools has to be fought for in each case; and in each example, the schools would definitely have preferred to do without it. But no principal had any difficulty understanding and appreciating the logic of the technique, and most principals had little difficulty living with its consequences over periods from four to six years. (But not all of them. As noted earlier, some replacement principals insisted on implementing their own school reforms that required eliminating their predecessor’s work that was under evaluation.)
The role of political will in implementing randomization is very important. In the health sciences, random assignment is common because it is institutionally supported by funding agencies and publishing outlets and is culturally supported through graduate training programs and the broadly accepted practice of clinical trials. The health-related studies done in schools tap into this same institutional and cultural structure. Something similar is also true of the rapidly growing number of studies of pre-school education that use random assignment. Most are the product of several forces: congressional requirements to assign at random; the high political and scholarly visibility of the Perry Pre-School (Schweinhart, Barnes & Weikart, 1993) and Abecedarian projects (Ramey & Campbell, 1991) that used random assignment; and the involvement of researchers trained in psychology and micro-economics where random assignment is valued. Agriculture is another field with a funding and training regimen that favors random assignment, even in schools (St. Pierre, Cook & Straw, 1981; Connell, Turner & Mason, 1985). So, too, are marketing and research on survey research.

Contrast this with education writ large. Reports from the Office of Educational Research and Improvement (OERI) are supposed to detail what is known to work. But neither the work of Vinovskis (1998) nor my own haphazard reading of OERI reports suggests any privilege being accorded to random assignment. Moreover, one recent report I read on bi-lingual education repeated old saws about the impossibility of doing such studies in education and claimed that alternatives are available that are as good—in this case quasi-experiments. In addition, at a recent foundation meeting on Teaching and Learning a representative of nine regional governors spoke about lists of best practices that are being widely disseminated. He did not care, and he believed governors do not care, about the technical quality of the designs generating these lists; the major concern is that educators can deliver to political actors a consensus on each practice. When asked how many of these best practices depended on randomized experiments, he guessed it would be close to zero. Several nationally known educational researchers were also present. They too replied that random assignments probably played no role in generating these best practice lists. No one present seemed to feel any distress at this.
I surmise there is little will to implement random assignment in education--not out of ignorance--but out of the sense there is little opportunity to conduct such studies and little need for them, given the availability of less noxious alternatives. So, there is no infrastructure or intellectual culture supporting random assignment either in Schools of Education, or in federal and state Offices of Education, or in foundations employing graduates of Schools of Education. So long as the beliefs are widespread that random assignment cannot be implemented in much of education and cannot be maintained in those rare cases where it is implemented, there can never be the kind of pan-support for random assignment that is available in the research worlds concerned with health, agriculture, the military, pre-schools, health-in-schools, marketing and the improvement of survey research.

What are the conditions most conducive to being able to randomize? These include: when the treatment is of shorter duration; when no extensive retraining of teachers is required; when new patterns of coordination among school staff are minimal; when the demand for an innovation outstrips the supply; and when students are the unit of assignment (or perhaps classrooms but not whole schools). Our guess, therefore, is that it would be more feasible to study different curricula at random; to introduce new technologies at random; to give students tuition rebates for Catholic schools at random; to assign more or different forms of homework at random (or by classroom); to assign teachers trained in different ways to classes at random, etc. None of these studies would be easy; but all should be feasible so long as there is a will to make the random assignment work over time; so long as there is knowledge of all that we have learned over the last 20 years about how to implement and maintain such assignment; and so long as strong fall-back options are built into the design in case the random assignment breaks down. In my opinion, it will be very rare for the biases resulting from such a breakdown to be greater than those resulting from teacher, school or student self-selection into treatments. The limitations of statistical selection controls are less the smaller the initial selection bias and the better this selection has been directly observed (Holland, 1986).
Random assignment is premature because it assumes conditions that do not yet hold in education. Random assignment makes better sense when the intervention is based on strong substantive theory; when it occurs within well managed schools; when it is reasonable to assume that implementation quality will not vary much between the units implementing the change; and when any standardized implementation that is realized is faithful to program theory. These conditions are not often met, for schools are indeed large and complex social organizations where multiple programs are simultaneously ongoing and where management by principals and other administrators is often weak and caught up in day to day affairs (including building politics). Thus, the capacity to plan is not what it should be and, anyway, many principals do not know how to be effective instructional leaders. It is also striking how variably the purportedly same reform effort is often implemented across districts, schools, classrooms and students; and when several different educational models are contrasted how small the between-model variation is, being swamped by that between schools (Rivlin & Timpane, 1975; Stebbins et al, 1978). Standard implementation and theoretical fidelity to program guidelines cannot be taken for granted in complex schools where much coordination is required if the many different actors in a building are to get anything accomplished (Berman & McLaughlin, 1977).

As the research emphasis shifted in the 1970's to understanding schools as complex social organizations with severe management and implementation problems, randomized experiments must have seemed premature. A more pressing need was for to understand management and implementation, and to this end more and more political scientists and sociologists of organizations were recruited into Schools of Education. They brought with them their own strongly held preference for qualitative methods and their memories of the wars between quantitative and qualitative methods in their own disciplines. Though they would not conceptualize it quite this way, part of their agenda in education was to increase the feasibility of responsible reform and evaluation through developing improved theories of how to manage schools and raise the quality of program implementation. These two topics have continued to be two of the major foci of educational research, each premised on understanding...
School research need not be predicated only on schools as complex organizations. An earlier conceptualization of the school was as the physical structure containing the many self-contained classrooms in which teachers tried to deliver effective curricula using instructional practices that demonstrably enhance students’ academic performance. This approach privileged curriculum design and instructional practice, not the school-wide factors that have come to dominate within the framework of schools as complex organizations—viz., strong leadership, clear and supportive links to the world outside of school, creating a building-wide communitarian climate focused on learning, and engaging in multiple forms of professional development, not just those relevant to curriculum and teaching matters. Many important consequences have followed from the intellectual shift in how schools are conceptualized. One is the lesser profile accorded to curriculum and instructional practice and to what happens once the teacher closes the classroom door; another is the view that random assignment is premature, given its dependence on positive school management and quality program implementation; and another is that quantitative techniques have only marginal utility for understanding schools, since a school’s governance, culture and management are best understood through intensive case studies, often ethnographic.

It is a mistake to believe that random assignment requires either wellspecified program theories, or good management, or standardized treatment implementation, or treatments that are totally faithful to program theory, however desirable these four features are on other grounds. Experiments protect against bias in estimates, not imprecision. Still, such sources of variation do require randomized experiments to have larger sample sizes if they are to attain the same statistical power as in less variable and better understood settings. Without this, the greater imprecision inclines towards a no-difference finding. When this seems to be the case, Cronbach (1982) has noted that many researchers do not stop their analysis there. They go on to conduct internal analyses based on stratifying schools by the degree
to which they were faithful in implementing program particulars, relating this variation to variations in the
planned outcomes. This strategy makes any resulting causal claim the product of the very type of non-
experimental analysis whose weaknesses only random assignment can overcome. This dead-end
suggests we focus on four other things: (1) avoiding the need for such internal analyses by designing the
experiment so that the original sample sizes reflect the expected extraneous variation; (2) anticipating
some sources of variation and taking steps in the research design to reduce through design those that
can be reduced; (3) studying implementation quality as a dependent variable to ascertain which types of
schools and teachers implement an intervention better--a topic severely underplayed in traditional
experimental design texts but central to program effectiveness; and (4) using measurement and statistical
procedures to reduce the impact of expected sources of irrelevant variation. Variable implementation
has implications for budgets and sample sizes, but it does not by itself invalidate the utility of random
assignment.

The aim of experiments is not to explain all sources of variation; it is to probe whether the
school reform idea makes a difference at the margin, despite whatever variation occurs in schools,
teachers, students or other factors. It is not an argument against random assignment to claim that many
reform theories are under specified, some schools are chaotic, treatment implementation is highly
variable, and treatments are not completely theory-faithful. Random assignment does not have to be
postponed while we learn more about school management and implementation. However, the more we
know about these matters the better we can randomize, the more reliable effects are likely to be, and
the more management and implementation issues will be worthy objects of study within experiments. No
advocate of random assignment will be credible in educational circles who assumes treatment
homogeneity or setting invariance, and experimenters need to be up-front that school-level variation will
be very large and may even be greater than in other research fields. It is not absolutely clear that schools
are indeed more complex than some of the other settings where experiments are routinely done--say,
hospitals. But most school researchers believe this, and it seems like a reasonable and politic working
assumption.
Random assignment does not deserve any special privilege since it entails trade-offs not worth making. Random assignment places the priority on unbiased answers to descriptive causal questions. Few educational researchers share this priority, particularly those who believe that techniques for achieving such clear causal inferences usually compromise other research priorities. Thus, Cronbach (1982) has argued strongly against Campbell’s assertion (Campbell & Stanley, 1963) that internal validity is the sine qua non of experimentation, arguing instead that external validity deserves at least as high a priority.

Internal validity is about the plausibility of causal inferences and depends on the clarity with which a set of previously identified threats to causal inference have been ruled out. External validity is about the generalization of any causal claim across settings, persons, treatments, outcomes and times that may, or may not, be the target universes around which a research plan was originally constructed. For Cronbach, external validity concerns both generalizing to the universes specified in a research plan and extrapolating to universes not so specified.

One context where disagreement between Campbell and Cronbach is concrete concerns some of the tradeoffs often required for random assignment. Experimental studies are often limited in time and space, with nation-wide experiments being rare. In education, experiments are also often limited to schools willing to tolerate having no choice over the particular treatment they receive and willing to undergo whatever measurement burdens are involved in assessing treatment implementation, mediating processes and individual outcomes, whether they received the treatment they preferred or not. What kinds of schools will make themselves so available? Would it not be preferable, Cronbach asks, to have a broader and more representative population of schools even if this entails causal inferences with more uncertainty? Why, he asks, should experimenters value uncertainty reduction about cause so much that they use such a highly conservative statistical criterion for inferring effects (p < .05). In real life, if we were in serious need we would decide to adopt a potentially life-saving procedure using a much more liberal risk calculus than this. In the same vein, Cronbach asks why experimental traditions should be so strict that schools not implementing any of the treatment are included in the analysis as though
implementation were perfect, just because the intent was to treat them. He believes that this is just another example of a counterproductive conservative bias that protects against wrongly concluding that a treatment was effective at the cost of failing to detect true treatment effects that may not initially emerge with great clarity. He also worries about the purism of those experimental enthusiasts who will not explore the data for unplanned comparisons involving treatment interactions with student, teacher or school characteristics, even though for Cronbach the true world of causal relationships is more complexly ordered than some invariant main effect.

He adds to these stringent conventions the charge that unplanned variation in implementation is not a cause for shame, but an opportunity to explore the reasons for such variation and the consequences of it. He also notes how poorly some experimental questions are framed, and argues for escaping from this framing whenever new and more helpful questions emerge, even if this emergence is midway through a study and the new questions cannot be answered in unbiased fashion. Since many educational programs do change with time, non-causal questions can easily become more central than they were earlier and new causal questions can arise. Cronbach believes that many experiments take so long to plan, mount, run and analyze that answering the causal issues entombed within them often entail answering an antiquated question.

Another trade-off experiments force, in his opinion, is between the utility of two conceptions of cause. The most prized questions in science are not about the descriptive causal connections to which random assignment is addressed, but rather about generative causal processes like gravity, relativity, DNA, nuclear fusion, aspirin, ethnic and gender identity, infant attachment, school-based management or engaged time on task. These are all constructs that imply instantiating processes capable of bringing about important effects in a multitude of quite different settings and times. They are so much more general in their application than learning whether one way of organizing schools affected student achievement at one time point in the sample of schools that volunteered to be in an experiment. Like most other educational evaluators, Cronbach wants evaluations to explain why programs work and, to
this end, he is prepared to tolerate more uncertainty about whether they work. So, he rejects the stringent causal standards of most experimental traditions, believing that they obscure many other lessons worth learning about a reform, especially the identification of causal explanatory processes that might be transferable to novel settings. Cronbach wants evaluation to pursue the traditional scholarly goal of full explanation but not through the preferred methods of science, opining that the methods of the historian, journalist and ethnographer provide better models for learning what happened in a reform and why.

A final trade-off is worth mentioning. Experiments seek to maximize the truth about the consequences of a planned reform. Their intended audience is usually some policy-making group, though their achieved audience may be books and journals that serve as historical archives. Experiments rarely seek to maximize the utility to personnel in the sampled schools. To them, a summary about the utility of a reform is usually less helpful than feedback about how to improve management and implementation in their own local school without much disruption. So, utility is more important than truth; the information needs of local personnel are more valued than those of amorphous policy-makers; and the immediacy of information needs argues against waiting until final reports are completed. Instead, the researcher should be someone who is prepared to be an informant to the school at all times, using whatever evaluative information has emerged and whatever relevant background knowledge he or she might already have. To await formal study completion is to invite local irrelevance.

These criticisms remind us that experiments should not be conducted unless there is a clear causal question that has been widely probed for its presumed long-term utility to a wide range of policy actors. They also remind us that, while experiments optimize on causal descriptive questions, that need not preclude either examining the reasons for variation in implementation quality or seeking to identify some of the processes through which a treatment influences an effect. The criticisms further remind us that experiments do tend to be conservative, sometimes so preoccupied with bias protection that other concerns are secondary and under researched. But this need not be so. There is no compelling need for
such stringent alpha rates; only statistical convention is at play here, not statistical wisdom. Nor need one conduct only analyses based on intent-to-treat, though such analyses do need to be included among all those done. Nor need one close one’s eyes to all statistical interactions, so long as the probes are done with substantive theory and statistical power in mind, so long as internal and external replications are used to check against profligate error rates, and so long as statements about likely interactions are couched in a more tentative way than conclusions that directly result from random assignment.

Researchers can also try to replicate experimental results by means of non-experimental analyses conducted on representative samples, realizing that such analyses have little standing by themselves and provide only indirect guidelines about extrapolating to larger populations. Finally, many controlled experiments would be improved by collecting ethnographic data in all treatment groups so as to examine possible mediating processes and unintended outcomes as well as to provide continuous feedback for self-improvement to all the experimental and control schools alike. This last procedure restricts generalization; but with newer programs and those that would include a monitoring component when bought to scale, the trade-off might be worthwhile. Experiments need not be as rigid as they are portrayed in some of the more compulsive sources on clinical trials.

Even so, there are bounds that cannot be crossed. All the above suggestions involve adding to experiments, not replacing them. To conduct a study with randomly selected schools but no random assignment to treatments would be to gamble on achieving wide generalizability of what may not really be a dependable causal connection. Conversely, to embark on an experiment presupposes the cardinal utility of causal connections; otherwise one would not do such a study in the first place. To be more concrete, in education the utility of experiments depends on avoiding the costs of wrongly concluding that Catholic schools are superior to public ones in the inner city, that vouchers raise achievement, or that small schools are superior to larger ones. Controlled experiments protect against recommending changes that don’t work and against overlooking changes that do, though this last possibility supposes experiments with considerable statistical power. Power analyses should be routine in experimental work on schools. No justification for random assignment is more central than identifying valid causal
knowledge so as to promote actions with likely positive consequences, to protect against implementing ineffective reforms, and to reduce the odds that actions will be taken without any systematic causal knowledge.

It is now 30 years since vouchers were proposed, and we have no clear answer about them. It is 30 years since Comer began his work that has resulted in the School Development Program, and again we have no clear answer; it is almost 20 year since Levin began accelerated schools, and here too we have no answer. While premature experimentation is indeed a danger--because there is little point evaluating what may be theoretically muddled or not implementable by ordinary human beings--these time lines are inexcusable. The Obie-Porter legislation cites Comer’s program as a proven exemplar worth replicating elsewhere and provides funds for this. But as I have reviewed the evidence elsewhere (Cook et al, 1999), when the legislation passed the only available evidence about the program consisted of testimony, a dozen or so empirical studies by the program’s own staff that were conducted in different locales and used primitive quasi-experimental designs; and the most cited single study (Comer, 1988) confounded the court-ordered introduction of the program with a simultaneously ordered reduction in class sizes of 40%. From this research base a federal decision was made that Comer’s program is effective and worth sponsored dissemination. This may have been the best decision to make based on the evidence then available; but to be restricted to such evidence about a causal connection verges on the irresponsible. Comer’s program is not different from any other program in this regard; we use it only as an example and not because the state of the information describing its results is particularly nefarious. Sadly, it is not. The trade-off Cronbach is prepared to make favoring generalization over causal knowledge runs just the risk exemplified by the literature on Comer’s School Development Program. It fails to appreciate that experiments are not meant to be representative; they are meant to be the strongest possible tests of causal hypotheses.

Yet Cronbach is not “wrong”. Causal hypotheses are special if they have both withstood strong falsification attempts via controlled experiments and their results are also demonstrably generalizable (or
if the boundary conditions limiting generalization have been empirically specified). Generalizing causal connections is a real problem with individual experiments (Cook, 1993). Unlike in medicine or public health, there is no tradition in education even of multi-site experiments with national reach. Single experiments of unclear reach are what we typically find, done only in Milwaukee or Washington or Chicago or Tennessee. Moreover, with some kinds of school reform there is no fixed protocol, and it is possible to imagine implementing vouchers, charter schools or programs like Comer’s or Total Quality Management schools in many different ways. Indeed, the Comer programs in Prince George’s County, Chicago and Detroit are different from each other in many, major specifics, given how much latitude districts are supposed to have in how they define and implement the program. So, while it is possible to argue that experiments can be made larger and more heterogeneous in terms of location and types of school, the non-standardization of many treatments requires even larger samples than those typically used in medicine and public health. Getting cooperation from so many schools is not easy, given the history of local control in education and the absence of a tradition of random assignment. Still, larger individual experiments can be conducted than is the case today.

But that may not always be wise. A sampling approach to causal generalization emphasizes replication across populations of schools, students and teachers, across settings throughout the nation, across locally reinvented treatment variants, and across outcome measures. Such variation may be better achieved through programs of research with many smaller experiments than by enlarging the sampling plan of a single study. Certainly, the laboratory sciences have progressed through a tradition of heterogeneous replication in a few labs and then assuming causal generalization until later results force one to conclude otherwise. They can do this because many of their causal knowledge claims are routinely replicated later as procedures required for the next stage of research on some phenomenon. But in areas where experimental traditions are weak—as in education—there can be no substitute for learning by doing, for conducting smaller (but adequately statistically powered) experiments in a staggered fashion across sites rather than putting all one’s eggs into a single large study basket. Whatever the merits of larger single experiments or phased programs of experiments, the point is that
experiments do not have to be both small and stand-alone things. Single experiments will not produce
definitive answers to any causal question, and they certainly will not answer all the ancillary questions
about causal contingencies. Like all science, experiments need and deserve a cumulative context.

The heterogeneity of sampling details cannot guarantee causal generalization when it is
understood as identifying those causal generative processes that apply in nearly all types of schools and
in multiple leaning contexts, as with the role of engaged time-on-task in stimulating student achievement
through homework, summer classes, longer school days, more interesting curricula, etc. It is therefore
important that steps be taken to identify the processes explaining why a particular effect came about in
some experiment or program of experiments. Many methods are available for this. Some are linked to
the measurement and quantitative analysis of data collected about these presumed mediating processes,
and others to the use of historians or ethnographers (but probably not journalists!) to explore what
happened and why in each treatment of a study. Any knowledge about explanatory processes so
resulting will be tentative, but worthwhile as part of the system of information from which inferences are
eventually drawn about generalized mechanisms that bring about desired outcomes. Cronbach’s
advocacy of causal explanation over causal description does not reflect a genuine duality. We can and
should attempt to describe and explain causal relationships.

Random assignment is linked to a model of research utilization that is rarely valid. Experimentation seeks
to recreate a specific model of rational decision-making. This requires laying out the alternatives (the
treatments), then deciding on decision criteria (the outcomes), then collecting data on each criterion per
treatment (data collection), and finally making a decision (based on the nature and size of the change
observed and the utilities attached to each criterion). The implication is that following such a procedure
allows a rational policy choice to be made of the best alternative.

However, empirical examination of how social science data are used in policy formulation leads
to claims that such instrumental choice is very rare (Weiss & Bucuvalas, M.J., 1977; Weiss, 1988).
Instead, when social science is used, the information is often less systematic, based on a rather diffuse process of “enlightenment” that blends information from existing background theories, from personal testimony, broad extrapolations from surveys, the consensus of a field (however achieved), claims from “experts” who may or may not have interests to defend, and novel concepts that are au courant and broadly applied—like the urban underclass or social capital have recently been in sociology. No privilege is extended to science writ large in this conception; nor to experiments. Indeed, the claim is made that “instrumental” usage is rare, understanding this to mean using specific social science research findings as the input for deciding how to modify a policy or ameliorate a program. Yet “instrumental” use is just what the experiment is designed to bring about.

In this tradition, it is also worth noting that decisions are multiply determined, and that central roles are played by politics, personality, windows of opportunity and values. Given this, scientific information of the type experiments produce can at best play a marginal decision-making role. In addition, many decisions are not “made” in the systematic sense of that verb built into the experimental model. Rather, they are “slipped into” or they “accrete”, with earlier small decisions constraining later large ones. Finally, we should not forget that official decision-making bodies turn over in composition, with new persons and issues replacing older ones. When studies take a long time to complete—as with many experiments on whole school reform—results may not be available until the instantiating issue is no longer “hot” or even “lukewarm”. Why conduct experiments if many of them are destined to be verdicts in history books rather than inputs into debates about current educational reforms? The argument is that the real world of research utilization is more complex than the rational choice model describes; and it is rare to find the kind of instrumental use on which the utility of the model is predicated.

Critics also point to another complicating empirical fact about use. Experiments rarely provide uncontested verdicts on reforms. In the educational policy world they do not enjoy absolute privilege as sources of causal knowledge. Disputes typically arise about whether the causal question asked was framed as it should have been, whether the claimed results are valid, whether all relevant outcomes were
assessed, and whether the proffered recommendations follow from the results. All social science findings tend to meet with a disputed reception, if not about the quality of answers provided, then at least about the questions not addressed. The logical control over selection that makes experiments so valuable does not entail that they are seen as gold standards that put to rest all quibbles about the validity of causal claims. Consider in this regard the reexaminations of the Milwaukee voucher study and the very different conclusions offered about whether and where there were effects (Witte, 1998; Green, Peterson, Du, Boeger & Frazier, 1996). Consider, also, the Tennessee class size experiment and the different effect sizes that have been generated (Finn & Achilles, 1990; Mosteller, Light & Sachs, 1996; Hanushek, in press). At issue with these examples are real scholarly disagreements, while in other cases the dispute also reflects some contribution from stakeholders protecting their own financial and cognitive interests. Policy insiders use multiple criteria for making decisions, and scientific knowledge of causal influences is never uniquely determinative.

But claims are made that a policy was changed because of experimental results. Nave et al (1999) implied this for the Tennessee class size experiment. But closer examination shows that those results did not exist in a vacuum. They were in line with what a much-cited meta-analysis had already described (Glass & Smith, 1979); they are consonant with theories that say children gain more if they are engaged and on-task with school work; they conform with teachers’ hopes, desires and expectations; they are in line with parents’ commonsense notions of facilitating children’s learning; and the results came at a time when the governor of Tennessee had national political ambitions, was using education as an individuating policy priority, had the state resources to increase educational investments, and he knew his actions would be popular both with the teachers unions (mostly Democrats) and with business interests in his own Republican party. So, his use of the experimental results cannot be ascribed to random assignment alone, though closer analysis may show it played a facilitative role at the margin.

Experiments exist on a smaller scale than would pertain if the services they test were to become state- or nation-wide. This scale issue is serious (Elmore, 1996). Consider what happened when the
class size results were implemented state-wide in Tennessee and California. This new policy was begun at a time of a growing national teacher shortage due to demographic shifts. Therefore the policy may have led to teacher-poaching across and within states and districts. Presumably, the richer states and districts would have recruited more and better teachers. Also, more new teachers were needed, and this involved some individuals leaving their old professions and jobs in order to become teachers. Were these new entrants equal in quality to existing teachers? Problems also arose in creating the greater numbers of classrooms required for class sizes to decrease, leading to the greater use of trailers and dilapidated buildings. Are the benefits expected from smaller classes to some extent undercut by the worse physical plant, the new hires from outside of education, and the redistribution of quality teachers? To go from experimental results to broad policy can change program implementation dynamics considerably, making those in the more local experiment different from the full scale policy implementation. The argument is that it is sometimes dangerous to use results generated by small scale controlled experiments that cannot mimic in all details what would happen if a policy were implemented on a broader scale.

There is some substance to these objections about the fit between the theory of use undergirding randomized experiments and the ways in which social science data are actually used. But the objections are exaggerated. Instrumental use does occur (Chelimski, 1987), and more often than the very low base rates readers might infer from most of the research denigrating instrumental usage. Moreover, one reason why some study results are widely disseminated is probably because random assignment confers credibility on them, at least in some quarters. This certainly happened with the Tennessee class size study and the pre-school studies cited earlier. Moreover, some decisions are made in what is close to a classical decision-theoretic sense. Witness the Obie-Porter legislation that listed educational programs thought to be demonstrably implementable and effective. Surely this legislation would have benefited from better input on the causal consequences of the preferred programs, had this been available? We should also not think it is trivial to create an archive of results that includes some studies whose results were “cold” before they came in. Some policy ideas get recycled and appear on later agendas--as
happened with vouchers; and other ideas enter texts used in graduate schools to train the next generation of professionals in a field (Leviton & Cook, 1983).

It is true that many experiments cannot deal with the scale issue; but then there was no compelling reason why Tennessee and California had to implement state-wide at one time. Could they not have phased-in the introduction of their new policy in ways closer to the experiment’s scale, using annual increases in the number of schools covered to explore scale issues? And although it is also true that experimental results enter policy debates as contested inputs, the din may be less deafening with a program of experimental studies than when reliance is placed on a single experiment. The goal should clearly be reliance on programs of experiments rather than individual stand-alone ones. In addition, conflicts about information are--and should be--endemic to a democratic political process. We should be wary of putting too much faith in any one method, since methods tend to have function-specific and not general strengths. Hence, population surveys are better for describing populations than for inferring causal relationships within them; in contrast, experiments are better for inferring causal relationships than for generalizing these relationships widely. The key issue, therefore, is whether doing an experiment reduces the volume of criticism about the validity of a causal claim. Criticisms about question formulation, sampling design, measurement choice, substantive interpretation of the results, etc. do not speak to the defining strength of controlled experiments. They are only relevant as criticisms of the experiment to the extent that conducting an experiment caused these other limitations in the study design. This has definitely happened in the past; but close attention to the trade-offs involved should minimize future problems without entirely eliminating them.

There is no necessary trade-off between instrumental and enlightenment usage. It is preposterous to think that experiments do not contribute to general enlightenment--about the kinds of interventions most and least implementable; about the influence principal turnover has on management; about the utility of theories that have no explicit component dealing with what happens once the teacher closes her classroom door; about the kinds of principals who are most attracted to school-based
management theories; about the kinds of teachers most amenable to professional development, etc. The era of black box experiments is long past. We need to learn about the determinants and consequences of implementation quality; it is legitimate to want to describe and measure constructs from the substantive theory undergirding a program; there is room for collecting qualitative as well as quantitative data so long as the data collection protocol is identical in all treatment groups; there are ways of getting stakeholder groups involved in the formulation and revision of guiding experimental questions; there are ways to get multiple stakeholders involved in interpreting the substantive relevance of experimental findings. All these procedures help generate enlightenment as well as instrumental usage. There is no necessary dichotomy between the two, just as there is not for some other dichotomies rampant in education research--e.g., between the need to use only qualitative or quantitative methods (Cook & Reichardt, 1979) or between positivist and post-positivist science (Cook, 1985).

Random assignment is not needed because there are other less noxious methods for generating causal knowledge. It is an old truism that no social science method will die, whatever its imperfections, unless a demonstrably better or simpler method is available to replace it. Most researchers who evaluate educational reforms believe there are superior alternatives to the randomized experiment, and so they are willing to let it wither and die. These methods are superior, they believe, because they are more acceptable to school personnel, the knowledge they generate reduces enough uncertainty about causation to be useful, the knowledge is relevant to a broader array of important issues than just identifying a casual connection, and schools are especially likely to be use the results for self-improvement. No single alternative is universally recommended, and here we discuss only three: intensive qualitative case studies, theory-based evaluations and quasi-experiments.

1. **Intensive Case Studies.** The call for intensive case studies came not only from people trained outside of the quantitative social sciences (e.g. Scriven, 1976) but also from some scholars who had first made their name as quantitative researchers and had then switched to qualitative case methods out of a disillusionment with “positivist” science and with the results achieved in the early round of educational
evaluations (e.g., Guba; Stake; House). Even Cronbach (1982) came to assert that the appropriate methods for educational evaluation are those of the historian, journalist and ethnographer, pointedly not the scientist. In any event, among evaluators within education the majority now seem to prefer case study methods for learning about reforms.

Converts often have a special zeal and capacity to convince, and when former proponents of measurement and classical experimental design were seen attacking what they had once espoused this may have had a strong influence on young students of methods in education. To be sure, some psychometricians within education refused to be converted, and the more elite graduate schools of education hired some scholars trained in the quantitative disciplines of statistics, economics and psychology. But none of these groups worked hard in print to advocate for experiments--Boruch and Murnane excepted--and none of them personally conducted educational evaluations in ways that would have demonstrated their feasibility and utility. Instead, some of them worked on higher education rather than primary and secondary schools and some of them worked on topics tangential to causal inference--e.g., developing newer psychometric techniques, reconceptualizing individual change and its measurement, developing hierarchical models with school, student and intra-individual change factors, and improving the practice of meta-analysis. So, when advocates of case methods wrote about qualitative evaluation techniques in education, this was in a context where there were few, if any, systematic rejoinders from organized countervoices. To add to the sense of solidity, many of their conclusions about research methods overlapped with those of scholars with doctorates in sociology and political science who had entered education with the conviction that schools should be understood as complex organizations best researched using the qualitative methods of organizational sociology and empirical policy analysis. So, two sets of voices converged to denigrate education research as a traditional science.

A central belief among advocates of qualitative methods is that these methods alone are capable of simultaneously giving feedback on the many different kinds of issues worth raising about a reform--
issues about the quality of problem formulation, the theory of the program, the quality of implementation, the determinants of such implementation, the proximal and distal effects a reform achieves, the unanticipated side effects that come about, the subgroups of teachers and students who are more and less affected by the reform, the other factors co-determining when and why effects appear, the relevance of the findings to various stakeholder groups and their fit to existing studies and policy concerns. Flexibility in the many types of insights offered became a salient attribute of qualitative evaluations, a flexibility the randomized experiment cannot easily match given its central function of facilitating clear causal inferences.

Advocates of qualitative methods also contend that they are successful in reducing some uncertainty about causal connections. Cronbach (1982) agrees that this might be less than in an experiment, but he opines that journalists, historians and ethnographers regularly learn the truth about causal connections just as many lay persons do in their own lives. Reflection, observation, observation-influenced reflection on the first reflections, more observation to reflect on the revised reflections, and so on. This is the iterative strategy of hypothesis generation and revision that theorists of ethnography have long advocated for testing causal hypotheses (e.g., Becker, 1970), and it is a strategy containing most of the elements of verification and falsification found in philosophy of science texts. The results so achieved can be richer than those built into the traditional black box experiment. This is because ethnography requires attention to the unfolding of explanatory processes at different stages in a program’s implementation, generating details about process at all times in a program's development that can be fed back to school personnel and used to explain why a program seems to be, say, effective in some areas but not others.

I do not doubt that some uncertainty reduction can be achieved through non-experimental empirical methods and hard thought. I do not doubt that these procedures sometimes reduce all reasonable uncertainty, though it will be difficult to know when this has been achieved. However, I do doubt whether intensive, qualitative case studies can reduce as much uncertainty about cause as a true
experiment. This is because such intensive case methods rarely involve a totally credible causal counterfactual. Since they typically do not involve use of comparison groups, it is difficult to know how the group under study would have changed over time without the reform under analysis. If control groups are added, unless they are randomly created it will not be clear whether the two groups would have changed over time at comparable rates. But note that the argument is about relative success in reducing uncertainty about a cause-effect connection. Whether these intensive case methods reduce enough uncertainty to be generally useful is a poorly specified proposition it is difficult to answer well. Yet this possibility does force us to note that experiments are justified only when a case can be made that a very high standard of certainty is required about a causal claim.

Two other things are more clear. First, schools are less squeamish about allowing in ethnographers than experimentalists (under the assumption that the ethnographers are not in classes and staff meetings every day); and second, feeding ongoing evaluation results back to teachers and principals with whom a skillful ethnographer has an ongoing relationship is especially likely to generate use of the data collected. Such use is highly local, within a single school, and so less grandiose than the usual aspiration for experiments--to guide policy changes that will affect large numbers of districts and schools. But in the work of educational evaluators like Stake, Guba and Lincoln, such local use is a desideratum, given how unsure they are that policy dictates issued from central authorities will ever be complied with once the classroom door closes.

2. Theory-Based Evaluations. It is currently fashionable in many foundation and some scholarly circles to espouse a theory of evaluation for complex social settings like communities and schools that does not depend on random assignment (Connell et al, 1995). Rather it depends on three steps: (1) 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 some major distal outcome, like achievement gains; (2) measuring each of the constructs specified in the substantive theory; and (3) analyzing the data to assess the extent to which the postulated relationships have actually occurred through time. For
shorter time periods, the data analysis will involve only the first part of a postulated casual model; but for longer periods the whole model might be involved. In this conception of evaluation, the burden rests on highly specific substantive theory, high quality measurement at each stage in the model, and valid data analyses of multivariate causal explanatory processes through time.

What makes theory-based evaluation an alternative to random assignment are several assumptions. The first is that it is not necessary to construct a valid causal counterfactual by randomly creating treatment and comparison groups. This theory-based approach has no necessary link to control groups; a study can be restricted only to groups experiencing the treatment under evaluation. Another is that demonstrating data patterns congruent with program theory indicates the validity of the theory. So, even in cases where only a third of the temporal links have taken place, say, the assumption following this is that the rest of the postulated process is more likely to occur and bring about the prized distal outcomes. Conversely, disconfirming the initial part of the reform theory suggests that the reform is not likely to be effective in the longer term. Advantages of corroborating any part of the theory is that the results can be used to inform program staff, to argue for maintaining the program, to provide a rationale for acting as though the program were effective, and to defend against premature summative evaluations that declare a program ineffective before sufficient time has elapsed for all the processes to occur that are presumed necessary for ultimate change.

Few will argue against a greater use of substantive theory in evaluation. The sole exceptions are probably those advocates of black-box work who believe that measuring theoretical processes is inappropriate in those circumstances where bringing an experiment to scale would not include process measurement. However, experiments can easily accommodate more process measurement, and they are improved thereby. It then becomes possible to probe, first, whether the intervention could have caused the theoretically specified intervening processes and second, whether these processes could plausibly have caused distal outcomes of interest. The first of these tests will be unbiased because it examines whether each step in the causal model is related to the planned treatment contrast. But the
second test will be biased if it depends on stratifying units by the extent to which the postulated theoretical processes are faithfully reproduced as a prelude to examining how this implementation variation is related to variations in some outcome. Still, I argue that such observational analyses are worth doing, though their results should be clearly labeled as more tentative than the results of any planned experimental contrast. There is little debate about the utility of measuring theoretically postulated processes in experiments and including them in the data analysis. The issue is whether such theory-based evaluation can function as an alternative to random assignment.

I am skeptical. 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. Moreover, it could be made more 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 entire flow of influence. It is all a little bit too neat for our more 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 it is difficult to know whether the next step in the model has not occurred yet or will not occur at all. Fourth, the method places a great premium on knowing, not just when to measure, but how to measure. Failure to corroborate the model could therefore be the product of the only partial validity of measures rather than the validity of the theory per se. Researchers can protect against this, of course, with more reliable measures or more multi-method measurement; but all this, while desirable, is burdensome on staff and students. Fifth there is the epistemological problem that many different models can usually be fit to any single pattern of data, and the causal modeling methods espoused do not permit falsifying among competing models. Thus, theory-based evaluations are predicated on prediction and not explanation, all appearances to the contrary (Glymour, Sprites & Scheines, 1987).
But the biggest problem is the absence of a valid counterfactual, knowing what would have happened had there not been a treatment. As a result, it is logically impossible to say whether the processes that occur are a genuine product of the intervention or whether they would have occurred without any reform. One way to guard against this is to have “signed causes” (Scriven, 1976), a multivariate pattern of relationships among variables that is so unique it could not have occurred for any reasons other than the availability of the reform. But signed causes depend on the presence of much well-validated substantive theory (Cook & Campbell, 1979). So, a much better safeguard is to have at least one comparison group, and the best comparison group is a randomly constructed one. So, we are back again with random assignment.

Do not get me wrong on theory-based evaluation. I am in favor of it when used as an adjunct to experiments; but I am against it in two contexts. The first is when it is suggested as an alternative to random assignment. The second is when it is used to postpone doing experiments because the most-valued policy outcomes seem so conceptually and temporally distant from the intended reform particulars. The developers’ and funders’ hope is always that implementing a reform with vigor and theoretical fidelity will entail little dilution of influence across the whole chain of probabilistic links in the theory, assuming the theory to be correct. But when implementation is weaker, the planned intervening processes may not come about even if the theory is true. So it is tempting to concentrate on the first theoretical steps without using control groups rather than to gamble on obtaining distal effects in experimental tests where these effects can only occur if the intervening steps actually occur and if they do so with sufficient strength that they impact on the next link in the hypothesized chain of influence. Theory-based evaluation could easily become an excuse not to evaluate by summative criteria.

3. Quasi-Experiments. The vast majority of educational evaluators favor intensive case studies, and a few of them plus some outsiders from psychology are now espousing theory-based methods. There are other education scholars, though, who deny the feasibility of random assignment and prefer quasi-experimental methods. These are mostly researchers interested in substantive topics and who evaluate
propositions about educational effectiveness with no intent to further evaluation theory and practice. Rather, they want to learn what is effective in their particular sub-field. We have already commented about how the existing evaluations of Comer’s program were of this type; the same is true of what we know about bilingual education and school desegregation, for example.

Quasi-experiments are like controlled experiments in purpose and in most details of structure. The defining difference is the absence of random assignment and hence of a demonstrably valid causal counterfactual. The essence of quasi-experimentation is the search, more through design than statistical adjustment techniques, to create the best possible approximation or approximations to this missing counterfactual. To this end, there are invocations (Corrin & Cook, 1998; Shadish & Cook, in press) to create stably matched comparison groups; to use age or sibling controls; to measure behavior at several points in time before a treatment begins so as to better estimate possible differences in pre-treatment trends; to extend the pre-treatment measurement of observations to create a time-series of observations; to look out for situations where units are assigned to treatment solely because of their score on some scale— as with draft numbers in the Vietnam War period, college grades for going onto the Dean’s List, or reported income for eligibility for various government programs; to assign the same treatment to different stably matched groups at different times, so that they can alternate in their functions as treatment and control groups; to build multiple outcome variables into studies, some of which should be influenced by a treatment and others not, provided that it is reasonable to assume that the latter will be influenced by the most plausible alternative interpretations. These are the most important elements from which quasi-experimental designs are created through a mixing process that tailors the design to the problem and resources available.

However good they are, quasi-experiments are second-best to randomized experiments when it comes to the clarity of causal conclusions. In some quarters, the word has come to connote any study that is not an experiment, or any study that seeks to build in some type of non-equivalent control group or pre-treatment observations. The word is rarely used in the intended sense of generating the best
possible approximation to a true experiment. Indeed, many of the studies called “quasi-experiments” are incompetent, including those on educational reform. Little thought is given to the quality of the match to create control groups in those cases where there are controls; to the possibility of multiple hypothesis tests rather than a single one; to the possibility of getting several pre-treatment time points rather than a single one--if there is even one; to the possibility of getting several comparison groups per treatment, one initially outperforming the treatment group and the other underperforming it to create a bracketed set of controls, etc. Reading quasi-experimental studies of educational reform projects often makes me feel ashamed of having contributed in any minor way to the institutionalization of that term, so weak are the designs and so primitive are the statistical analyses of those designs that really nothing could put right. Although I am convinced that the best quasi-experiments give “reasonable” approximations to the results of randomized experiments, to judge by the quality of the work I know best--on school desegregation, Comer’s School Development Program and bi-lingual education--the average quasi-experiment in these fields is lamentable with respect to the confidence it inspires in causal conclusions. Recent advances in the design and analysis of quasi-experiments are not getting into educational research where they are sorely needed. Nonetheless, it is telling that the best estimate of the validity of any quasi-experiment is to compare it with a randomized experiment on the same topic. It is always the fall-back and not the preferred option; and all quasi-experimental designs are definitely not equal.

Conclusions

1. Many independent sources indicate that random assignment is very rare in educational research concerned with the effectiveness of reforms designed to improve primary and secondary school academic performance. Yet many research issues in education are of the very form for which controlled experimentation was uniquely developed--viz, can something deliberately manipulated change a valued outcome?
2. Controlled experiments are nonetheless common in schools when the aim is to learn about strategies to prevent mental health problems, violence, drug use or even unhealthy nutritional practices. It is not clear why experiments are more frequent in these domains than on more traditional academic reform issues. Individual and collective political will may be one explanation, since most of the school-based prevention researchers were trained in public health and psychology where random assignment is held in high esteem. Moreover, random assignment is valued by the funding sources and journal editors to whom prevention researchers address their work. Capacity may be another explanation. Most school-based prevention experiments are of short duration; they involve curriculum interventions; the implementers are usually researchers and not teachers; and the topics selected may engage educators (and parents) less than issues of school governance and classroom teaching. More research is needed on why this disciplinary difference occurs in the incidence of school-based random assignment. It should suggest many of the conditions promoting such assignment.

3. An intellectual culture exists within the research establishment concerned with evaluating educational reforms and understanding school management that is characterized by multiple beliefs, any combination of which could sustain the (erroneous) conviction that randomized experiments are not worthwhile. These beliefs include: Experiments are philosophically naive in the theory of causation they espouse and in the assumptions about causal orderings they make. Experiments are not very practical, being rarely feasible in schools and rarely implemented well in the rare cases they are begun. Experiments require trade-offs with other methods and lower the quality of answers to important questions about school reform, including questions about management, the determinants and consequences of variation in implementation quality, and the identification of causal explanatory processes. Also, the information experiments generate about cause is of a type that is rarely used to change educational policies. And, anyway, the same information could be gained by other means that are less noxious to school staff and more flexible in the range of questions addressed. In such an intellectual culture it is not surprising that randomized experiments are rare and not valued.
4. Many of the beliefs above are poorly warranted, and I provide reasons for this throughout the text—especially as regards beliefs about the availability of less noxious and more flexible alternatives. None of them provides as convincing a causal counterfactual as the randomized experiment. However, other criticisms are not unreasonable and have important implications for proposing practical adjuncts both to the bare bones structure of experiments and to their monolithic emphasis on identifying descriptive causal connections. It is especially important to describe implementation quality, to relate implementation to outcome changes, to describe program theory, to reliably measure the extent to which theoretically specified intervening processes have actually occurred, and to relate these processes to outcomes. Efforts should also be made to increase the chances that an experiment poses a simple and clear causal question whose importance is widely recognized and to increase the number of experiments done on any one topic.

5. It will be difficult to persuade the current community of educational researchers to begin doing randomized experiments solely by informing them about the advantages of this technique, by providing them with lists of successfully completed experiments, by telling them about new methods for implementing randomization, by exposing them to critiques of the alternative methods they prefer, and by having prestigious persons and institutions outside of education recommend to them that experiments be done. Although I have not had much time to describe the social organization of the research community concerned with evaluating educational reforms, it is a community in which all parties share at least some of the beliefs outlined above. Hence, many members are convinced that anyone pursuing a scientific model of knowledge growth is an out-of-date positivist seeking to resuscitate debates that are rightly dead. So the community sees little value in better connections to recent research design as understood in statistics or psychology.

6. Some rapprochement might be possible, though the extent of this is not at all clear to me. At a minimum, such rapprochement requires advocates of experimentation like myself to be explicit about the real limits of the technique, to engage our critics in open dialog about their objections to randomization,
and to show how experiments are improved by greater sensitivity to issues they value that relate to program theory, implementation specifics, quantitative and qualitative data collection, attention to causal contingency, and concern with the management needs of school personnel as well as central decision makers. My prediction is that unless these steps are taken, it will be difficult to enlist the current community of researchers on educational reform behind any banner promoting increased use of randomized experiments. It will also be difficult with those mid-level government and foundation officials trained in educational research methods over the last 20 years. All these groups share beliefs about the inappropriateness of a science model of knowledge growth and about the need for an R & D model with links elsewhere than to clinical trials.

7. Though it is desirable to enlist the current community of education evaluation specialists, it is not necessary to do so. They are not part of the tiny flurry of controlled experimentation now occurring in schools. Moreover, in several substantive areas Congress has shown its willingness to mandate that controlled studies be done, especially in early childhood education and job training. So, "end runs" around the educational research community are conceivable, implying that future experiments could be carried out by contract research firms like MDRC, Abt or Mathematica, or by university faculty with a policy science background, or by educational faculty who are now lying fallow. It would be a shame if such a strategy lowered access to those researchers most knowledgeable about micro-level school processes, school management, how school reforms are actually implemented, and how school, state and federal officials tend to use educational research. It would be counterproductive if outsiders to school reform research had to learn anew the craft knowledge insiders already enjoy. Such knowledge genuinely complements controlled experiments; it is not in any necessary intellectual opposition to it.

8. We advocates of random assignment have to be careful lest our enthusiasm fail to acknowledge the real difficulties with the technique and lead us to be seen as naive proponents of gold standards made of clay.
9. I am an empirical social scientist knee-deep in the trenches of quantification and hence neither a student of cultural analysis nor a historian of ideas. So, by myself, I have no hope of recognizing any of the important assumptions I may have violated in this paper. If it is true that “No out-of-discipline experience goes unpunished,” then I will have to pay the usual heavy price genuine scholarship extracts from dilettantes.

References


Case study/case series: A set of case reports that describe some observations in a small number of patients (persons). These frequently lead to the... Documents Similar To Critical Appraisal of Education Research. Carousel Previous Carousel Next. The Basic Idea Underlying Experimental Research is to Assess the Impact Of. Uploaded by Che Soh. Performance appraisals were viewed in much the same way as tests; that is to say, they were evaluated against criteria of validity, reliability, and freedom from bias. The emphasis throughout was on reducing rating errors, which was assumed to improve the accuracy of measurement. The research addressed two issues almost exclusively—the nature and quality of the scales to be used to assess performance and rater training. Performance appraisals are different from the typical standardized test in that the "test" in this case is a combination of the scale and the person who completes the rating. And, contrary to standardized test administration, the context in which the appraisal process takes place is difficult if not impossible to standardize. Crombie, The Pocket Guide to Critical Appraisal; the critical appraisal approach used by the Oxford Centre for Evidence Medicine, checklists of the Dutch Cochrane Centre, BMJ editor's checklists and the checklists of the EPPI Centre. Case study research Research methods in familiy therapy. 393-410. There is no level A evidence to determine open repair (OR) or endovascular repair (ER) intervention to treat popliteal artery aneurysms. This article addresses a case of endovascular treatment of both popliteal artery aneurysm (PAA) and below knee peripheral atherosclerotic occlusive disease (PAD) in which we opted for endovascular treatment with satisfactory outcome, as well as a discussion about indications and expected results.