Obstacles to and Limitations of Social Experiments: 15 False Alarms

by
Stephen H. Bell
Laura R. Peck

May 22, 2012

manuscript submitted to
International Conference on Field Experiments in Policy Evaluation

AUTHOR INFORMATION:
Stephen H. Bell, Ph.D., Principal Scientist & Abt Fellow, Abt Associates Inc., Social & Economic Policy Division, 4550 Montgomery Avenue, Suite 800N, Bethesda, MD 20814 USA

Laura R. Peck, Ph.D., Principal Scientist, Abt Associates Inc., Social & Economic Policy Division, 4550 Montgomery Avenue, Suite 800N, Bethesda, MD 20814 USA
Laura_Peck@abtassoc.com / (301) 347-0557 tel / (301) 634-1801 fax

KEY WORDS:
evaluation design, social experiments, internal validity, external validity

Obstacles to and Limitations of Social Experiments: 15 False Alarms

ABSTRACT

When deciding how to allocate limited taxpayer or donor funds for social programs, policy makers and program managers increasingly ask for evidence of effectiveness based on studies that do not raise quibbles over methodology. They want to *know* the extent to which their interventions have their intended effects based on research that all sides of the policy debate can agree provides credible scientific evidence. The basic claim for the "social experiment"—that the "coin flip" of randomization creates two statistically equivalent groups that cannot subsequently move apart except through the effects of a successful intervention—makes resulting estimates unbiased measures of the intervention's impact. Despite the transparency and conceptual strength of the experimental strategy for revealing the causal link between intervention and impact, experiments are often criticized on a variety of factors. We address 15 of these concerns in this paper and find each of them less objectionable than is widely believed. Among these, we believe reluctance to bear the higher costs of social experiments compared to non-experimental methods to be short-sighted, given that the cost of not knowing a program's impacts—or lack thereof—can be much greater than the costs of finding out. Moreover, an assortment of issues concerning ethics, scientific integrity, and practical feasibility need not stand in the way of expanded use of social experiments by governments and foundation funders seeking to convincingly and accurately measure the impacts of their social policies and programs.

Obstacles to and Limitations of Social Experiments:
15 False Alarms

When deciding how to allocate limited taxpayer or donor funds for social programs, policy makers and program managers increasingly ask for evidence of effectiveness based on studies that do not raise quibbles over methodology. They want to know the extent to which programs have their intended effects on their target populations based on research that all sides of the policy debate can agree provides credible scientific evidence. If it is true, as Trochim notes, that "Only a few programs *should* survive in the long run" (2009, 28), then it is our contention that decisions on termination, continuation, or expansion should be based on research meeting high standards of evidence. For government and foundation policy makers, strong causal inference showing that a public sector or philanthropic intervention has a favorable impact provides justification for continued funding or expansion. Likewise, unequivocal evidence is often needed to justify termination of an existing program with strong political or bureaucratic constituency but that, with rigorous testing, is found to be producing little or no social benefit.

An abundance of research methods seek to answer the "Does this intervention work?" question in social policy. This multiplicity of strategies for deriving evidence of effectiveness springs mostly from the need to construct an estimate of counterfactual—i.e., what *would have happened* to program participants in the absence of the intervention—in order to gauge the extent to which actual outcomes represent an improvement over the status quo. It is only by *increasing* the well-being of the target population above what it would have been otherwise that a policy intervention can be justified, because the counterfactual level of well-being could always be attained without such an investment.

It is not our intent to debate the relative merits of the various designs or approaches that evaluators use to measure program impacts. We recognize that there are other important policy questions besides those regarding impacts on the target population.[1] Instead, we recognize—as many others have—that implementing a classically-designed experiment, which randomly assigns potential program participants to “treatment” and “control” groups, provides one promising method for producing a valid counterfactual (the control group) and hence making accurate causal inference. Randomized experiments are widely heralded for ruling out other plausible explanations for favorable outcomes that occur, rival explanations that are legion in our complex socio-economic environment suffused with many other policy actions and cultural forces brought to bear on program targets. The argument goes that, other than through impacts truly caused by the program being randomized, it is only by chance that the outcomes of treatment and control group members can differ. Furthermore, in large enough samples chance too can be ruled out, leaving program impact as the only remaining explanation for outcome differentials.

Despite the transparency and conceptual strength of the experimental design in establishing a causal link between intervention and impact, experiments are often criticized on a variety of factors. This paper addresses 15 of these criticisms and finds each criticism less objectionable than is widely believed. We classify the objections into four categories: ethical, scientific, practical, and financial. Analysis of each concern leads to the overall conclusion that it need not threaten the reliability or applicability of experimental techniques for measuring the impacts of social interventions. As a result, we believe government agencies and foundation

[1] One of the most widely used evaluation textbooks – Rossi, Lipsey & Freeman’s *Evaluation: A Systematic Approach*, now in its seventh edition (2004) – suggests a hierarchy considerations that should come into play in considering a program’s merit (p.80): (1) need for the program, as concerns a social ill that should be addressed; (2) persuasiveness of the program design in theory and based on past evidence; (3) success of the program’s process and implementation; (4) program outcomes/impact; and (5) program cost and efficiency in achieving those results. The first three steps in the hierarchy are not about program *impact* but about program motivation and operations, and for these purposes the question of a valid counterfactual does not arise. Our paper centers on the fourth goal of the hierarchy—achieving program impacts—and on one particular methodology for measuring that achievement—a random assignment experiment.

funders have the opportunity to use experimental methods to obtain transparent and compelling answers to important social policy impact questions in many more instances than they may recognize.

WHAT ARE SOCIAL EXPERIMENTS?

We begin by ensuring that the experimental methodology for measuring social program impacts is understood, apart from the objections to be addressed later. Most people unfamiliar with the concept of randomized social experiments find experience from the medical field a useful introduction: in order to test whether a new drug is effective in its claims, pharmaceutical companies undertake “randomized control trials” (RCTs). These trials randomly assign some people to take the new drug while others take a placebo, an inert dose. By following subjects’ subsequent outcomes, researchers can determine not only the extent to which the drug made a difference (in reducing headaches or ulcers or cancer), but also the extent to which side-effects accrue. Because the two groups are *randomly* assigned in their treatment experience, the only difference between the two is the medication.

Substitute “public policy,” “social program” or “intervention” of some sort for “drug” and the same approach can be taken to evaluate the effectiveness of public and non-profit efforts to ameliorate all sorts of social and economic ills. Social experiments deliberately exclude from participation in an intervention some of the people or organizations the intervention would ordinarily serve in order to create a control group that represents the world without that intervention. Excluded cases are selected from would-be participants purely by chance, through a lottery-like process that randomly divides the population into two groups: a “treatment group,” assigned to receive the program or policy that defines the intervention; and "control group," excluded from the program or policy for research purposes.

When truly selected at random from the potential participant pool and kept out of the intervention, the members of an experimental control group will meet three critical conditions for accurately representing the world without the policy/program. First, except by chance, they are collectively the same kind of people or organizations as the people or organizations in the treatment group. Second, they are not subject to the intervention, and therefore experience no effects from it. Third, they otherwise operate in entirely the same environment—policy, economic, and social—as the program participants in the treatment group, and therefore represent a true "counterfactual," what would have happened in the absence of the intervention.

In a successfully implemented experiment, this second condition assures that the control group differs from the treatment group on the factor of interest—the intervention whose impact we wish to measure—while the other two conditions assure that *nothing else between the two groups differs.* In large samples, with many cases allocated to the treatment or control groups on a purely random basis, any chance differences in preexisting characteristics (both measured and unmeasured) between the two groups tend to disappear, and it becomes very unlikely that observed differences in later outcomes between the two groups are caused by anything other than the effects of the program or policy under study.[2]

In short, reliably representing the world without the intervention is essential if government and philanthropic social programs are to be judged by the difference they make. Experiments that use random assignment—if successfully implemented and effective at meeting the challenges discussed later in the paper—provide a solid counterfactual as represented in the outcomes observed for the control group. This counterfactual allows elimination of the so-called

[2] Orr (1999) writes in depth about these and other properties of random assignment evaluations of social programs.

"threats to internal validity," or the plausible rival explanations for why change might occur over time or why differences could arise between participants and nonparticipant comparison groups determined by determined by natural processes (e.g., Campbell & Stanley, 1968; Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002). The plausible rival explanations, in scientific lingo, include systematic selection into the intervention, maturation, regression-to-the-mean, history, testing, and instrumentation.[3] The experimental design itself is structured to net out any of these influences in producing unbiased impact estimates.

Recognizing that they are not a panacea for all policy evaluation needs, we believe it is important that social experiments be conducted as widely as possible for the same reason that experiments are ubiquitous and invaluable in advancing our knowledge about chemistry, biology, medicine, agriculture, and industrial processes: to vary the one factor of paramount interest (in this case, a particular public policy or program) while holding all other factors equal. Policy makers and program administrators use information about impacts to decide on programs' future course. That said, experiments face a variety of challenges, which lead us to pose this paper's research questions as follows: What are the criticisms of primary social experiments? To what extent are those criticisms valid? To what extent are the issues raised by valid criticisms surmountable? The viewpoints expressed in our examination of these questions—which reflect years of designing and analyzing both random assignment and non-experimental impact evaluations in a variety of program contexts (e.g., employment and training, education, housing, family and child assistance, public assistance, and food and nutrition policy)—are intended to spur dialogue among policy evaluation researchers and funders and thus push the field forward.

[3] Interactions between these threats exist as well. It is beyond the scope of this paper to restate the fundamentals of various evaluation designs' strengths and weaknesses in addressing these threats to internal validity.

THE ETHICAL CONCERN

Despite the basic strength of the random assignment approach, social experiments have their limitations. We begin with the most fundamental challenge—the ethical concern that exclusion of some eligible and deserving individuals from a program's services or a policy's provisions for the sake of research is unethical.

Criticism #1: It's not ethical to have a control group.

An often-cited obstacle in planning an evaluation is concern about the ethics of randomizing access to government services (e.g., Boruch, 1997; Boruch, Victor & Cecil, 2000; Cook & Payne, 2002; Gueron, 2002). Are the individuals who "lose" the government "lottery" and enter the control group disadvantaged unfairly or unethically? Likewise, some programs are entitlements such that denying access for the sake of research is not only unethical but also illegal. Of course no evaluation should ever propose any illegal treatment of potential research subjects, but we question whether—in the case where we do not know whether a treatment might benefit participants at all—denying access is truly unethical. While this issue seems always to surface (and should), the fact remains that social experiments have been used often in the U.S. to evaluate the effectiveness of pilot and demonstration projects. Greenberg et al. (2004) catalogue over 200 of them. So at some level social experiments are ethically acceptable. Critics' main concern is that randomizing people to a control group denies them access to opportunities that could potentially benefit them. Three responses are possible.

First, if a program has to limit the total number of people or organizations served due to funding or administrative capacity constraints, *it will in some way ration access*. Random assignment, with control group members left out of the program's services, is just one way to ration. Whether it is a better or worse way is the real question. We argue that giving all deserving applicants an equal chance at access, through a lottery, is the *fairest, most ethical way* to ration (e.g., Bickman & Reich, 2009; Orr, 1999). Less fair is to allow program staff to choose their favorites, based on personality or some perception of whether staff believe (unscientifically) that the person would benefit from participating. In situations where programs have waiting lists, perhaps the fairest way to determine who should be next to participate is by a coin toss, rather than by imposing some arbitrary judgment regarding who should be next.

Second, if a program's effectiveness has yet to be determined, *being turned away from participation at part of a control group cannot be presumed to be worse than being admitted.* For example, if job training on average does not lead to better employment outcomes—the very question an impact study seeks to answer—then participating in it at best constitutes a neutral situation and may be a disadvantageous, at least for the time it wastes. One example of this is the Job Training Partnership Act. A large-scale randomized impact evaluation found this program to cause unemployed youths admitted into the program to wait longer to go to work than their counterparts in the experimental control group (Orr et al., 1996), possibly because they expected to program to deliver them an unrealistically attractive job, which did not happen. This example illustrates how the assumption that control group members will be harmed by exclusion from an untested social program runs counter to a fundamental research paradigm in many fields—that when studying interventions to see what they affect, we must presume *no* impacts until proven otherwise.[4] There is no basis for asserting as fact that randomly assigning someone

[4] Peter Rossi's (1987) "Iron Law of Evaluation" is that "The expected value of any net impact assessment of any large scale social program is zero" (p.4). Further, given that the zinc law is "only those programs that are likely to fail are evaluated" (p.5), we would be smart to start from the position that control group members are unlikely to be harmed.

to a treatment group will benefit her or him—or to make the logically equivalent assertion that assigning someone to a control group will *hurt* that person—until *after* an evaluation is conducted that demonstrates that this is the case. Of course, one should not exclude a control group from a beneficial intervention for research purposes a second time once proof of a benefit is in hand.   But believing in or hoping for a benefit is not the same thing as proof, and knowing that harm rather than benefit could also arise, a seemingly black and white ethical principle becomes ambiguous.

Third, it is possible that control group members will be disadvantaged, but there is justification for why this might still be an ethical decision to allow. Society, which benefits from good information on program effectiveness, may be justified in allowing small numbers of individuals or organizations to be disadvantaged in order to gather that information. Any potential disadvantage will be temporary, but benefits may be long-term. The interests of many future program participants, and every taxpayer, may legitimately outweigh the costs borne by a comparatively small number of control group members. In medical research, scientists deny treatment to a control group only until they learn that there is benefit from a previously unproven medication; once the medication is deemed effective, a much broader population can benefit. The consequences of denying access to a social program or policy are arguably much less dire than the potential life-or-death consequences in medical research, but we collectively still approve that type of research for the larger potential health benefits it can provide to society. This is a tricky ethical issue—but a potentially powerful argument—worthy of further attention.

No researchers can appropriately weigh the balance of these considerations in seeking justification for the broad use of randomized impact studies as a way to improve policy. That is a decision that must fall to public officials. Yet certainly, on the basis of these considerations, the question of the inherent fairness or lack of fairness of random assignment in social experiments remains open for debate, with arguments to be made on both sides of the issue. At the very least it argues that social experiments should not be unequivocally dismissed on the basis of ethics.

SCIENTIFIC CONCERNS

We next explore five alleged scientific limitations of social experiments to gauge each concern's validity and potential for remediation through effective research design. These criticisms arise from doing experiments in real-world conditions and may apply to evaluations of existing, ongoing programs or of pilot tests of new interventions.

Criticism #2: Experiments measure effects on those assigned to the treatment group, not necessarily on those who actually get the treatment.

The most easily dispensed with criticism of random assignment impact evaluations is the charge that they reveal only the impact of the "intention to treat"—called ITT impacts by Heckman et al. (2000)—rather than the impact of actually being treated—what Heckman et al. call the "impact of the treatment on the treated," or the TOT impact. This charge arises whenever less than 100 percent of the treatment group participates in the intervention, when the "treated" group is different (smaller) than the entire randomly assigned treatment group. Less than 100 percent participation is common in random assignment evaluations, since individuals cannot be compelled to take part in an intervention such as job training simply because they applied for it and appeared likely to participate at the point when they were randomized and offered admission.

On the one hand, the ITT estimator might be considered the most policy relevant estimate of a program's impacts: if new program design or options are being tested and the new features would not become mandatory, then understanding targets' overall response—including participation and subsequent changes in outcomes—tells policy-makers what they need to know. What happens if we change the public assistance benefit structure to be more generous for those who secure court ordered child support? (e.g., Hamilton, et al., 1996). The ITT-estimated impact captures the behavioral response to the new policy as well as any change in outcomes that occurred. The ITT essentially averages impacts across those who took up and did not take up the offer of treatment and represents what is likely to happen if such an intervention were rolled out still allowing for less than 100 percent "compliance."

On the other hand, the TOT estimator might still be of interest, and incomplete participation by the treatment group is also readily correctable. Applying what is called the "no-show adjustment" converts an ITT to a TOT. Formalized in the evaluation literature by Bloom (1984), this adjustment *assumes that the intervention has no effect on those members of the treatment group who never participate in any intervention activities*—for example, those randomly assigned to a voluntary after school program who never attend the program. This assumption is, in our experience, viewed as innocuous by almost all evaluators and policy analysts in most situations where it arises.

Where the assumption holds, the initial measure of impact—the intervention's average impact across all treatment group members, including both potentially positive (or negative) effects on participants and zero effects on non-participants (the "no-shows")—can be rescaled to the average impact *on just those who do participate* (i.e., the effect of the treatment on just "the treated”).[5] No assumptions regarding the similarity of participants and no-shows, or the ability of statistical methods to adjust for differences between them, are needed here; they can be as different as night and day and the result is still valid as long as the intervention has no effect on the no-shows.

In brief, the criticism that experiments examine impacts averaged across a treatment group comprised of takers and no-shows should not mean that we should not use experiments for exploring the impact of social policy innovations. Most all evaluations consider whether the ITT is the more appropriate impact estimate, and in situations where it is not, they apply the widely-accepted (and reasonable) Bloom correction.

Criticism #3: Experiments fail to compare an intervention’s services to no services at all, instead comparing the intervention to “everything else that’s out there.”

In a decentralized, fragmented federalist system, the policies and services of one branch of the national government will often be supplied in similar if not identical form by other government or nonprofit agencies. That is, unlike in medical trials where a “placebo” is intended to represent nothing,[6] social experiments’ “counterfactual” is usually described as the “*status quo*” or as “business as usual.” Random assignment does not control whether individuals access these “substitute” services; as a result, some control group members inevitably will do so. This means that the control group is not a “no services” placebo. This is the case, for example, when state pre-kindergarten programs do substantially the same things for members of the same target group as the federal Head Start program.

[5] The rescaling divides the original impact estimate by the participation rate in the treatment group (or, equivalently, by 1 minus the “no-show” rate), as explained by Bloom (1984). The result of this calculation gives a magnitude for the average effect of the intervention on *participants* necessary for the observed magnitude of the overall average treatment effect, including zero impact on “no-shows,” to occur.

[6] Nothing except, perhaps, for a person’s natural tendency to get better anyway (and perhaps get better faster with a good attitude).

This circumstance gives randomized impact studies the same character as the real-world programs they are intended to evaluate, and hence it is a strength rather than a weakness of the experimental approach. Just as is true of an experimental control group, some of the people who apply to the U.S. Department of Health and Human Services (DHHS) Head Start program would obtain similar assistance from other state sources were DHHS's intervention not available. Some other families would not—again, as is true of an experimental control group. It is precisely the choice between these two scenarios—a set of children enrolled in Head Start, or a subset of the same children served by state pre-kindergarten programs—that DHHS controls when implementing its Head Start program. If Head Start were not there, services to some of the children it serves would still take place, from other sources. DHHS should not seek to impose any stronger contrasts between the intervention and control group children when measuring the program's impact.[7] Knowing how a given intervention strategy, uniformly imposed on all members of the target group, compares to no intervention at all does not help social decision-making in a fragmented federalist system of many intervention sponsors and selective consumer participation.

Looking at "our services" compared to "everything else that's out there" is exactly what DHHS should be doing to justify its program and policy portfolio, because if "everything else that's out there" is enough, then the money spent on DHHS programs could be cut back without consequence. The research goal in this example is not to determine whether developmentally focused programs for disadvantaged preschoolers have value compared to a world where no such programs are available. Rather, it is to learn how much difference DHHS's involvement in this policy arena makes given all else that exists. Were DHHS not offering its particular

[7] If worried about the ability to statistically detect differences between treatment and control group outcomes when control group members receive alternative services, a larger study should be undertaken to obtain better—and adequate—statistical power.

intervention, some of those served would get something similar elsewhere, and for those families the value added by DHHS's program is truly lessened by the existence of alternatives.

Just as criticism #2 seems unjustified (the ITT estimate is usually of greater policy interest; and if not we can easily compute the TOT estimate), criticism #3 is also unjustified: the comparison of a treatment group's outcomes to a counterfactual represented by everything else that is out there provides the policy-relevant information needed to inform decisions about policy continuation, modification or termination.

Criticism #4: Counterfactual experiences in the control group are distorted by easier access to similar services from other sources than by the one being evaluated.

This concern—which we have encountered in our unpublished work but have not introduced previously to the literature—is about "queuing effects" for control group members. The argument articulated in criticism #3 for "letting happen what will happen" in the control group leads to a difficult but crucial question: Doesn't the *existence* of the intervention under study *shorten the queue for control group members seeking to access alternative services*? This occurs by putting treatment group members (and other, non-research individuals) into the program under study—something that would not happen if that program under study did not exist. Thus, the "competition" for alternative services is not as fierce as it would be, and the control group gets too much help from those sources.

To see how this could happen, consider the example of job training provided by programs funded by the U.S. Department of Labor (DOL) and programs sponsored by other agencies. As a thought experiment suppose one of these programs, DOL's Workforce Investment Act (WIA) program, was completely eliminated in a given year. In this scenario, the total supply of services and the number of service slots would fall precipitously for the consumer groups served by the program. Where those people turn, and to what extent they access alternative services to help build employment skills, will strongly determine the importance of having WIA in place as it now exists compared to no WIA training at all.

If this is the policy choice Congress or the Department faces—continuing versus eliminating WIA-funded training—one would want to run an experiment in which (1) treatment group members are given access to WIA, and (2) control group members compete for access to training services from non-WIA sources, but they do so in a market in which treatment group members are also vying for those same alternative training slots. Unfortunately, the second condition cannot be met: the treatment group cannot simultaneously participate in the WIA treatment and jostle with the control group for access to the limited number of alternatives to WIA, sometimes squeezing them out of those slots.[8] Without the latter, we will see *too much use of alternative services in the control group*, and hence (if services improve outcomes) too small a difference in average outcomes between the treatment and control groups when measuring impacts. Controls do not accurately reflect the world without WIA, since in a true counterfactual world they would have to share non-WIA training slots with members of the treatment group as well as everyone else with whom they actually *do* share those slots.

The negative judgment just delivered on the reliability of the control group counterfactual hinges on two as yet unstated assumptions: impact evaluation results will guide a decision to either keep WIA at its current scale or eliminate it altogether; and other programs that provide similar services to the same consumer group *would not expand their scale were WIA eliminated.* If choosing between full-scale WIA and no WIA, and if expecting the "hole" it would leave to not be filled in at all by other employment and training service funders, then DOL would indeed

[8] This is what Rubin (e.g., 1974, 2010) refers to as SUTVA, the Stable Unit Treatment Value Assumption. For optimal experimental conditions treatment and control cases are assumed to experience distinct and un-interfering treatments. But, if the presence of one treatment affects the nature of the control condition, then SUTVA is violated.

want control group members to have to compete with treatment group members for other, non-WIA training slots to achieve the appropriate counterfactual for the policy choice it faces. But if DOL expects other funders to expand services in response to the “shortages” created by WIA’s disappearance, then added service availability for the control group up to a point represents the correct counterfactual. Other programs might expand their scale to make up most or all of the difference, at which point a sample of individuals participating in WIA contrasted with a sample participating in other similar programs would accurately depict the consequences of DOL’s policy decision.

A similar point holds when policy makers seek guidance for deciding whether to expand or contract WIA funding *at the margin.* Only a small share of all those seeking WIA-type services would be affected by a WIA capacity expansion or contraction at the margin. In this circumstance, what happens to control group members should represent well the options and outcomes of individuals who are marginally displaced from WIA—they really would not have to compete with the workers staying in that program as its size changes only fractionally. Thus, the contrast produced by the treatment-control comparison in an experiment would again trace the right consequences of DOL’s policy choice.

With most evaluations of existing programs likely to influence funding and scale at the margin rather than in an “all or nothing” way, and with the potential for partially offsetting adjustments in the scale of alternative services in a fragmented federal system, randomized experiments with *full* access to alternative services among control group members seems a better approximation to the desired evaluation counterfactual than experiments with *no* control group access. Neither is perfect, but in principle the perfect version of control group experiences is unknowable until policies are changed—either marginally or dramatically—and other agencies react—either a little or a lot. Absent that information, a cautious approach featuring marginal changes and more modest treatment-control differences in service access—i.e., the approach that most social experiments actually produce—provides the safer basis for policy assessment.

Criticism #5: Treatment group experiences are distorted by changes in program scale or changes in the population served.

Another intractable, but possibly minor, problem of randomized impact studies arises on the treatment group side for interventions with a fixed number of service slots when some of the people or organizations that would ordinarily occupy those slots are placed in a control group. Removing a portion of the normally-served population necessarily results in one of two changes to an existing program's operations: it serves fewer people, operating below capacity (or, if below capacity anyway, operating even further below capacity than usual); and/or it serves additional people who ordinarily would not be served due to capacity constraints. At the very least the program probably has to recruit more than it had been previously, which changes something about program operations regardless of the results on the characteristics of the population served. There is no way around this issue—if one artificially pulls out some would-be participants, one necessarily leaves the program short (or shorter than usual) of participants or bring in others who normally would not participate.

The question is whether either of these results matters to the size of the program impacts measured, the quantity one seeks to determine through random assignment? Likely both situations do matter, though perhaps not to a very great extent. A program with unnaturally created vacancies may deliver services differently for the customers it does serve. If budgets remain unchanged, then the typical participant in a less fully subscribed program may receive more services and experience a larger impact. Or lower numbers may change the dynamic of any group elements of the intervention that depend on how many participants interact in the service delivery setting (e.g., class size in educational interventions), either increasing or possibly diminishing impacts on those who participate in smaller groups.

Alternatively, program scale and operations could remain unchanged if added people are served that normally would be closed out due to capacity limits. These are clients of lower priority in the program's view, or clients with less motivation or ability to ensure that they make the first cut. In a normal year, when random exclusions are not imposed on those "ahead of them in line" for the sake of the research, they would not be served. Unless a lottery of some sort is *ordinarily* used to ration slots among a surplus set of applicants, the usual means of obtaining access *creates distinctions between those who get in and those who do not*—one of the very problems that experiments are used to overcome. It may be that the applicants thought most in need of help receive priority or that those expected to benefit most from the program's services (which might or might not be the same people) do. On one factor or another, then, entrants differ from the interested non-entrants, and these differences may correlate with the size of program impacts. This creates a selection problem of another sort, though one less likely to matter appreciably to the size of measured impacts: the intervention (treatment) and counterfactual (control) samples match one another through random assignment *but they both represent slightly the wrong set of people, a somewhat different and larger set than would ordinarily be served.*

In this situation, Olsen et al. (2007) propose a way to identify which individuals would ordinarily have been served so that impact results can be produced for just that subset. Local program operators can be given the opportunity to identify the applicants they would have enrolled in a normal year (i.e., a year without a lottery). Their incentive to do so reliably is an increased probability of those cases being randomly assigned to the treatment group rather than the control group. This set of participants and their control group counterparts can be analyzed as a subgroup defined by pre-random assignment information and the study can obtain impact estimates for the normally-served population.

No good data exist on how much these factors could matter to the size of impacts measured from the experimental data. What we do know is that both these problems—artificial shortfalls in enrollment and different-than-usual participant populations—diminish as the control group shrinks in size relative to the program's capacity. When control group members are spread over many local programs, with only one or two individual control group cases in any community, no program can be pushed much below its regular scale or forced to serve very many new customers by the removal into control status of some it normally would serve. The National Job Corps Study provides an excellent example of steering clear of distortions in the treatment group through minimal control group exclusions in any one local program site, while still achieving a large total control group and evaluation sample through inclusion of many sites (Schochet, et al., 2001). This model should be emulated whenever possible.

Criticism #6: Experiments eliminate selection bias only for the difference in policy exposure controlled by random assignment and not in other places where important impact questions arise (and might encounter selection bias when answered by the data).

A final scientific objection is that, while eliminating selection bias in measuring the effect of participation *at the point of random assignment*, experiments do not provide equally unequivocal information on the consequences of participation at other stages of the intake process. For example, they still leave evaluators with non-experimental comparisons of the relative impacts of different sequences or "dosages" of services that are only determined following randomization (and hence are never known for the control group).[9] Conversely, an experiment cannot show directly how much difference interaction with the program *prior to random assignment* might have made to participant outcomes, since these effects occur for both the treatment and control group members.

This emphasizes the importance of choosing wisely where to position random assignment in the intake flow and early steps of participation. But does this criticism create insurmountable liabilities for the experimental approach? No, since other types of impact analysis are equally hobbled at all points where an experiment does *not* place random assignment *and are much more hobbled where the experiment does place random assignment.* Solving one selection bias problem—the one that could distort answers to the most important policy question to be addressed by a study—is clearly a virtue of experiments over solving none.

FEASIBILITY CONCERNS

Beginning with the alleged ethical and scientific failings of random assignment impact evaluations, we have argued that these criticisms are not the terminus for experiments but instead only issues to consider in building stronger impact evaluations. The next step would seem straightforward: "Just do them." But it is not. Feasibility issues regarding the kinds of situations in which social experiments can be utilized—which we now discuss—also demand consideration. Our conclusion in this realm is that all eight of the identified feasibility concerns can be overcome with adequate resources devoted to the evaluation, if the policy question to be addressed is of sufficient importance. Naturally, the policy questions of insufficient importance

[9] Nevertheless, some techniques exist that capitalize on random assignment when defining and analyzing subgroups defined by post-random assignment events, choices or program features (e.g., Gibson, 2003; Kemple & Snipes, 1999; Morris & Hendra, 2007; Peck, 2003,2005, 2007; Schochet & Burkhardt, 2007). These techniques reduce the security of criticism #6 by providing opportunities to explore questions of dosage or varying treatment paths within the experimental design.

to justify spending capable of redressing feasibility challenges should not be studied with experimental methods, unless all other more feasible methods are equally expensive (see discussion of funding tradeoffs at Criticism #15 below).

Criticism #7: Saturation interventions that affect entire local communities can not be randomly assigned.

It has become commonplace for evaluations of systems change and other community-wide "saturation" interventions to be evaluated non-experimentally (e.g., Connell, Kubish, Schorr, & Weiss, 1998; Fulbright-Anderson, Kubish, & Connell, 2002). Even the most sophisticated (and expensive) of those evaluations face substantial challenges in providing support for causal claims, connecting intervention and any observed changes in outcomes. We posit that the entire endeavor of evaluating community-wide change efforts would be a prime candidate for an experimental design: The U.S. is a very large nation, with thousands of local communities or neighborhoods that could be randomly assigned into or out of a particular policy or intervention. Saturation intervention also makes data collection more difficult and expensive and any impacts that do occur harder to find if diffused across many people in the community, but these drawbacks afflict any impact analysis of saturation interventions, not just experiments. At the very least, *that* an intervention involves community saturation is not a sufficient argument to dismiss using an experimental design to evaluate it impacts.

Criticism #8: Programs that struggle to meet enrollment targets should not be subjected to random assignment.

Concern understandably arises in situations where programs are already struggling to enroll targets. One possibility is that programs with this experience are not in demand and perhaps should not be continued, in lieu of those for which there is greater demand. On the other hand, to include these programs as part of an experimental evaluation would provide the needed causal evidence to make that decision. In that case, only a few control group cases need to be sampled in any locality, as long as enough localities can be included in the study, in order to provide the information needed in an evaluation. Additional technical assistance resources can help raise application counts sufficiently to accommodate a modest-sized control group so that such a program could be included in an experimental evaluation.

Criticism #9: Random assignment is not appropriate for extremely long-term interventions whose consequences cannot be fully tested in experimental settings.

Some policy innovations seek to alter citizens' or firms' behavior in areas guided by long-term planning, such as the decision to return to work in the face of permanent disability or investment in new production plant capacity. We would not expect policies intended to affect such behaviors to reveal their full impact in an experimental setting unless the treatment group members in a study believe the policy will apply to them forever (or at least for many many years) and the control group members believe the policy will never apply to them. The concern is that these are unrealistic conditions to impose when testing a new policy intervention—i.e., that demonstration projects cannot create a credible sense of permanency for either group: control group members will come to expect their "turn" and behave in a way that anticipates the policy applying to them later on, and treatment group members will know that the demonstration is a test of something that may be withdrawn.

In response to this concern, we would make two points. First, government's treatment of its citizens and businesses changes all the time, making uncertainty about how long current policies will continue the right context for observing behavior under different current "rules" as is done in a random assignment experiment. Second, policy conditions for treatment and control group members in an evaluation do not have to be unnaturally abbreviated just because they are assigned at random. Control group "embargoes" from the tested intervention need not be time-limited unless ethical concerns become too extreme, and treatment group interventions can be offered and funded for a lifetime if that aspect of the alternative policy is sufficiently important to its success. One example is the provision of alternative disability benefits to individuals whose medical condition is not expected to improve. Changes in benefit rules designed to encourage work may have no effect, or less than their full effect unless treatment group members believe these changes will apply to them over their entire lifetimes. We do not see this as an obstacle to accurate experimental findings: it simply means that adequate funding needs to be committed to pay for changes to benefit rules for treatment group members that extend over their entire lifetimes. If initial findings of positive effects—given the long-run planning decisions of treatment group members—are sufficiently encouraging, then the intervention can be applied to the control group within their lifetimes; what matters in this case is that those control group members *were not expecting* this change to occur—certainly a reasonable assumption given the inertia of existing provisions in long-established social programs.

Criticism #10: Random assignment is not appropriate for interventions with low participation following randomization, because average effects for the treatment group as a whole will be too weak to detect.

Two responses to this criticism are warranted. First, small average effects can be detected in sufficiently large samples and readily translated into average effects among participants through the "no-show" adjustment discussed earlier. Second, also as noted earlier, the estimate of the intent-to-treat is often highly policy relevant. Particularly for programs that would remain voluntary, the ITT estimator is the appropriate one, even when some might find the TOT estimator interesting.

Criticism #11: Experiments do not inform questions of program effectiveness when interventions have multiple facets and the impacts of the individual facets are of interest in their own right.

It is common for government agencies commissioning policy evaluations to ask researchers to tell them whether a program has its desired impact overall and, if so, *what features of the program account for its effectiveness.* The latter information will point up the facets of the intervention whose "dosage" might be increased to obtain larger impacts, as well as the facets that could be eliminated (to reduce costs) without loss of effectiveness. The "up/down" nature of experimental findings concerning the intervention as a whole—when an experiment applies the entire intervention with all its facets to the treatment group and none of those facets to the control group is thought to severely limit the usefulness of random assignment as a way to discover how a program could be made *more* effective or less costly without loss of effectiveness.

The response to this criticism of evaluations based on random assignment is obvious: randomize more things. Recent examples such as the U.S. Department of Housing and Urban Development's Family Choices evaluation of assistance to homeless families have conducted random assignment to two or three different intervention models in addition to a control group in order to determine which of the policy features that differ between the different treatment models (e.g., duration of housing subsidy, provision of social services in addition to housing assistance) are essential to improving on the control condition. While "multi-arm" random assignment is unusual in recent social experiments (another example is the National Evaluation of Welfare to Work Strategies) it need not stay that way. Indeed "early social experiments were much more ambitious" in randomizing to multiple variants of an intervention (Bloom, 1995, p.18), with the Negative Income Tax (NIT) experiments of the 1970s, for example, randomly assigning families to as many as 58 distinct policy options by varying tax rates and guarantee levels to ascertain people's responses (Greenberg & Robins, 1986). While this approach may sacrifice statistical precision if not sufficiently funded (i.e., if fewer people are assigned to any one policy option), these examples highlight that lack of applicability to the question of "what works best" is not an inherent limitation of random assignment experiments.

Moreover, *multi-stage* random assignment can be used to answer questions about the effects of different treatment experiences without sacrificing statistical precision. For example, suppose a government agency wanted (as the United Kingdom government did ten years ago) to know if a work incentive would increase employment among people who receiving public income support benefits and whether the impacts of such incentives on employment success and self-sufficiency would be increased by providing those induced to go to work with additional supports such as transportation subsidies and case worker interdiction when on-the-job problems arise. Rather than randomize the target population into three groups at the outset—one treatment group receiving the added incentive, another treatment group receiving the added incentive plus work supports, and a control group—thereby cutting the size of each group by one-third relative to two-armed random assignment, an innovative design would randomize at two different points: two-way randomization of the full sample to determine which individuals receive the added work incentive, followed by—for individuals in the incentivized group who obtain jobs—separate two-way randomization of the provision or non-provision of the work supports. This design not only increases statistical precision of the impact estimates for a given total sample size by using some sample members for multiple purposes (e.g., the incentivized workers who do not obtain jobs serve to represent outcomes under an incentives-only policy and an incentives-plus-work-supports policy), it concentrates the examination of the impact of work supports on just individuals who, if assigned to that policy "package," actually use them, making impacts large enough for statistical detection more likely.[10]

In addition to randomizing across multiple treatment models or using multi-stage randomization to capture varying impacts in multi-faceted interventions, the analytic approaches described in response to criticism #3 apply here. Having randomly assigned targets to treatment and control status means that we have important variation that is exogenous to treatment assignment and can be used for advanced analyses of subgroups that follow varying treatment paths even without further randomization among different treatment regimens. In brief, this criticism is by no means one that should forestall the use of social experimentation; instead, there are viable alternatives in design and analysis that can help answer policy questions regarding multi-faceted interventions and heterogeneous impacts.

Criticism #12: Experiments do not capture the full effects of interventions that have "general equilibrium" consequences beyond the experimental sample.

In an interconnected world, some consequences of social policies can spill over to individuals not directly engaged in the program or services offered. This would happen, for example if job training equips workers to take jobs other workers would otherwise have held, resulting in earnings losses for workers perhaps not included in the research[11]—and certainly not included among the individuals randomly assigned at application to the program since workers compete for jobs with a broad swath of potential new hires. Economists call this kind of situation a "general equilibrium effect" because it ripples through the system creating impacts in secondary locations. All research based on data for those who directly participate in an intervention and a confined sample of non-participants such as an experimental control group

[10] This design was proposed to the U.K. government by the first author but not adopted.

[11] Smith (2000) provides several good examples of these general equilibrium effects of labor market interventions and explores the evaluation challenges they create.

faces this issue—spillover outside the research sample—not just randomized experimental designs. Indeed, while it does not capture these spillover effects directly, nor does randomization make them more difficult to measure. There may be instances in which one can use experimental data to place a useful upper or lower bound on the magnitude of potentially important general equilibrium effects as has been proposed for the Design Options for the Search for Employment project at the U.S. Department of Health and Human Services and is currently under development by the authors. More generally, general equilibrium analyses of social policy initiatives are always difficult, but no more so for having measured the direct effects of those policies experimentally.

Criticism #13: Experiments have limited generalizability, usually not being based on a statistically representative set of sites.

At least a half-dozen random assignment impact evaluations of ongoing social programs have now been conducted in geographically-based probability samples of the nation without substantial attrition of local programs from the research (Cook, DATE). The Food Stamp Employment and Training Evaluation (Puma et al., 1990) is one example, the National Head Start Impact Study (Puma et al., 2005) another. Scholars have debated the internal *v.* external validity tradeoff since the terms were invented (e.g., Bracht & Glass, 1986; Jimenez-Buedo & Miller, 2010), with a general consensus preferencing internal to external validity (Reichardt, 2011). But this does not put external validity out of reach of social experiments, as emphasized in recent work by Olsen et al. (2011) or Tipton and Hedges (2011), for example.

Criticism #14: Experiments take too long.

Some critics posit that policy decisions in which results are needed quickly—without a multi-year lag to set up and conduct random assignment and wait for medium- and long-term outcomes to emerge—cannot rely on experiments to inform them (e.g., Besharov, 2009).. We make several rebuttal points to this criticism. First, if policy-makers are interested in long-term outcomes and impacts, then any prospective evaluation design will require the time needed to cover the policy-relevant follow-up period. One possibility would be for government agencies to establish a system of regular and temporally-overlapping experimental evaluations of any ongoing program so that new information is always emerging from experimental data. In addition, evaluations of shorter-term outcomes need not take "too long." As Peck and Scott (2005) showed, a small, government intervention with six month follow-up took little more than six months to complete, informing policy decisions about modifying and expanding the innovation in a timely manner. Further, Ludwig, Kling and Mullainathan (2011) urge changing the outcomes of interest in experimental research to focus on the shorter-term mechanisms by which interventions have their effects, rather than the long term impacts that arise from some unknown causal chain.

THE FINANCIAL ISSUE

Criticism* #15: *Experiments are too expensive.

The eight technical criticisms raised above can generally be overcome with political will, sufficient funding, and competent evaluation management. That said, the financial costs of experiments, to those sponsoring the research and, hence, indirectly to taxpayers, have often been put forth as an important obstacle to their use (e.g., Orr, 1999). It is beyond the scope of this essay to investigate the costs of alternative methods in any detail. Suffice it to say that budgetary constraints on funding agencies—be they government, foundation or nonprofit—is not a valid reason to avoid experiments, especially in an era of heightened fiscal accountability and results-focused policy decision-making.

This is especially clear when one recognizes that the appropriate basis for choosing among competing research techniques is the *marginal* cost of experiments compared to other equally ambitious research studies that tackle the same set of policy questions. Obtaining broadly representative data on social program outcomes for thousands and thousands of people, both with and without a policy in place, is not inexpensive—a facet of cost invariant with *how* the program participants and non-participants are selected.

One exception is data from large surveys of households and workers collected for other non-evaluation purposes, such as the Current Population Survey and the Survey of Program Dynamics. In those instances the social cost of data collection has already been paid, and individual federal agencies can use the information at low cost. Unfortunately, reliance on national surveys of this sort to measure the impacts of social programs was the first non-experimental approach to impact evaluation to be discredited by careful methodological research (Barnow, 1987; LaLonde, 1986).

Another exception stems from the observation that small, local reforms provide an ideal testing ground for incremental changes that might be used to tweak policies and programs and subsequently improve performance. The Peck and Scott (2005) example cited earlier documents one state's efforts to change its public assistance intake process. Not a major reform, the innovation used an experimental design to ascertain the extent to which welfare recipients were any better off (in terms of their employment outcomes) when case workers used a more detailed intake assessment than they had previously been using. State program managers and analysts designed and implemented the intervention, provided data the on treatment and control cases' characteristics and outcomes, and university researchers analyzed the data to determine short-term impacts. The costs of this pilot test were marginal and the learning likely more definitive than it would have been were the State to simply have compared outcomes before and after a change in procedures. That is to say, not all social experiments are or need to be the large, national policy reforms that we have been discussing; and smaller, localized efforts can be affordable ways for program managers to learn more about how their changes in program operations are associated with changes in targets' outcomes.

On a larger scale, an important point about cost considers the "opportunity costs" of *failing* to do experiments—the money spent on ineffective programs that continue to be funded (and continue to offer false hope) because unbiased information on their inadequate impacts is unavailable. On this basis, some observers see the balance clearly swinging toward experiments as the comparatively *low cost* investment option compared to other methods once an appropriately broad social viewpoint is adopted. Not surprisingly, these have been among the most outspoken supporters of random assignment studies (e.g., Burtless & Orr, 1986; Orr, 1999).

Importantly, prominent researchers known for their contributions to non-experimental impact evaluation methods have more recently taken similar stances. Smith (2002), for example, writes: "Random assignment does have its costs, as it typically requires substantial staff training, ongoing staff monitoring and information provision to the potential participants… At the same time…this case can be overstated" (p.21).

Greenberg et al. (2004), in the Digest of Social Experiments, provide an important closing perspective: "Sponsoring a social experiment requires complex resource allocation decisions. The social experiments conducted to date were authorized by many different [individuals] representing a wide spectrum of political views… It is striking that many very different individuals decided that this type of investigation is worth its costs" (13). It would be difficult for today's national government to back away from this practice of rigorous, experimental impact evaluation on the grounds of insufficient funds, when reliable policy guidance is known to depend on the use of random assignment designs.

OTHER CRITICISMS

The 15 alleged obstacles that we have detailed pertain most directly to experimentally designed evaluations. There are also scientific and practical limitations that face large-scale policy impact evaluations of *any* design, experimental or non-experimental. These include incomplete data, limited sample sizes (especially when looking at effects on subgroups), inability to sort out causes of cross-site variation, and lack of assured reliability for national policy making when study sites are not nationally representative. Much has been made about these shortcomings in the literature questioning the appropriateness of experiments, often without acknowledgement that they are not unique to experiments. In fact, naturally occurring populations, one with and one without policy exposure, can be and often are studied using non-experimental methods (i) in non-representative locations, (ii) with incomplete data and (iii) little capability to sort out which subgroups benefit more or (iv) what accounts for differences in apparent impacts across subgroups and locales.

Another criticism of social experiments concerns incorrect analysis of nested or hierarchical data. This occurs when statistical modeling used to measure impacts ignores the level of the hierarchy at which randomization occurs. An example arises in research on education reform if entire schools are randomized in and out of the treatment but evaluators apply methods designed to produce reliable findings only if individual students are randomized (e.g., Bickman & Reich, 2009; Henry, 2009). We do not list this as a challenge for randomized evaluation designs to overcome; it is simply a problem of inappropriate analysis methods in what should be highly reliable, non-problematic social experiments. As the science of accurately analyzing nested data in an experimental context becomes more widespread (see for example Bloom, 2005, for an important contribution in this direction), we expect this issue to disappear.

DISCUSSION & CONCLUSIONS

The purpose of this paper is to identify the main critiques of experimental designs and explore the extent to which they these preclude the use of experiments in a range of social policy evaluation contexts. In terms of the primary, ethical criticism, we argue that a lottery is the fairest way to ration access to supply-limited social services. Doing so is justified even in situations without supply constraints when society does not know whether the "lottery winners" will fare better than the "losers" yet having that information is vital to making programmatic and funding decisions that help disadvantaged citizens going forward. The several scientific objections that have been raised about social experiments also appear to us to be unfounded or minor relative to the strength of the randomized evaluation approach and its potential to inform policy decisions about program effectiveness. The several practical obstacles to experiments' on closer inspection appear to be only "speed bumps" that do not impinge the road toward greater use of social experimentation. Finally, we conclude that the reluctance to bear the perceived higher cost of social experiments compared to non-experimental methods is short-sighted, given that the cost of not knowing about a program's causal effects may be potentially much greater than the cost of finding out.

We hope that our discussion of these potential pitfalls is useful to those in the government and foundation sectors as they plan for future evaluation activity. We also hope that some of the factors that may now be thought of as obstacles to using social experiments receive greater debate in the literature and other forms and perhaps as a result come to be seen as not so much of a challenge. A reassessment of the strengths and limitations of random assignment policy evaluations seems particularly appropriate in light of recent a advances in experimental evaluation design and analysis that have strengthened the ability of random assignment designs to address historical concerns in ways cited in this paper.

Are experiments sufficiently robust to serve as the customary standard of practice in impact evaluation? Operationally and scientifically, we believe our examination of the issues in this paper indicates that they are, particularly if the political will and funding commitments exist to carry them out properly. Should and will they be used more extensively? That depends in large measure on their costs compared to the costs of alternative research strategies, which deserve more careful inspection in specific applications of various research methods. In the meantime, the argument in this paper is that issues of ethics, scientific integrity, and feasibility need not stand in the way of expanded use of social experiments for measuring policy and program impacts. The commonly cited objections and limitations are, on closer inspection, in fact, false alarms.

REFERENCES

Bell, Stephen H. 2006. "Estimating Impacts of Social Experiments when Exposure to Treatment Drifts out of Alignment with Randomization Intent," paper presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Madison, WI.

Besharov, Doug. 2009. "From the Great Society Continuous Improvement Government: Shifting 'Does It Work?' to 'What Would Make It Better?'" *Journal of Policy Analysis and Management,* 28(2): 200-222.

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review,* 8 (April): 225-246.

Boruch, Robert F. 1997. Randomized experiments for planning and evaluation: A practical guide, Chapter 3. Thousand Oaks, CA: Sage Publications.

Boruch, Robert F., Timothy Victor and Joe S. Cecil. 2000. "Resolving Ethical and Legal Problems in Randomized Experiments." *Crime & Delinquency* 46(3), 330-353. DOI: 10.1177/0011128700046003005

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *The Journal of Human Resources,* 21(4): 606-639.

Bracht, Glenn H. and Gene V. Glass. 1968. "The External Validity of Experiments," *American Educational Research Journal*, 5(4): 437

Brickman, Leonard, and Stephanie M. Reich. 2009. "Randomized Controlled Trials: A Gold Standard with Feet of Clay?" in Donaldson, Stewart I., Christina A. Christie and Melvin M. Mark (eds). What Counts as Credible Evidence in Applied Research and Evaluation Practice?, pp. 51-77. Thousand Oaks, CA: Sage Publications, Inc.

Connell, James P. Connell, Anne C. Kubisch, Lisbeth B. Schorr, and Carol H. Weiss (eds.). 1998. New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts. Washington, DC: Aspen Institute.

Cook, Thomas D. and Monique R. Payne. 2002. "Objecting to the Objections to Using Random Assignment in Educational Research" in Evidence Matters: Randomized Trials in Education Research, Mosteller, Frederick and Robert F. Boruch, eds., Chapter 6, pp. 150-178.

Fulbright-Anderson, Karen, Anne C. Kubisch, and James P. Connell (eds.). 2002. New Approaches to Evaluating Community Initiatives, Vol. 2: Theory, Measurement, and Analysis. Washington, DC: Aspen Institute.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus experimental estimates of earnings impacts," *Annals of the American Academy of Political and Social Science*, 589, 63-93.

Gueron, Judith M. (2002) "The Politics of Random Assignment: Implementing Studies Affecting Policy" in Evidence Matters: Randomized Trials in Education Research, Mosteller, Frederick and Robert F. Boruch, eds., Chapter 2, pp. 15-49.

Greenberg, David, Mark Shroder, and Matthew Onstott. 2004. The Digest of Social Experiments, Third Edition. Washington, DC: The Urban Institute Press.

Hamilton, William L., Nancy R. Burstein, August J. Baker, Allison Earle, Stephanie Gluckman, Laura Peck, and Alan White. 1996. *The New York State Child Assistance Program: Five year impacts, costs, and benefits*. Cambridge, MA: Abt Associates Inc.

Jimenez-Buedo, Maria and Luis M. Miller. 2010. "Why a Trade-Off? The Relationship between the External and Internal Validity of Experiments," *Theoria*, 69: 301-321.

Kemple, James J., and Jason C Snipes. 2000. *Career academies: Impacts on students' engagement and performance in high school*. New York, NY: Manpower Demonstration Research Corporation.

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*, 25(3): 17-38.

Olsen, Robert, Stephen Bell, and Jeremy Luellen. 2007. "A Novel Design for Improving External Validity in Random Assignment Experiments," paper presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington, DC.

Olsen, Robert B., Larry L. Orr, Stephen H. Bell, and Elizabeth A. Stuart. 2011. "The Loss of External Validity in Policy Evaluations that Choose Sites Purposively: Bias in the Impact Estimates and Recommendations for Bias Reduction," Under review (revise and resubmit at the *Journal of Policy Analysis and Management*)

Orr, Larry L. 1999. Social Experiments: Evaluating Public Programs with Experimental Methods. Thousand Oaks, CA: Sage Publications.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. Does Training for the Disadvantaged Work? Evidence from the National JTPA Study. Washington, DC: The Urban Institute Press.

Peck, Laura R. 2003. "Subgroup analysis in social experiments: Measuring program impacts based on posttreatment choice," *American Journal of Evaluation*, 24(2): 157–187.

Peck, Laura R. 2005. "Using Cluster Analysis in Program Evaluation," *Evaluation Review*, 29(2): 178-196.

Peck, Laura R. 2007. "What are the Effects of Welfare Sanction Policies? Or, Using Propensity Scores as a Subgroup Indicator to Learn More from Social Experiments," *American Journal of Evaluation*, 28(3): 256-274.

Peck, Laura R. and Ronald J. Scott, Jr. 2005. "Can Welfare Case Management Increase Employment? Evidence from a Pilot Program Evaluation," *Policy Studies Journal*, 33(4), 509-533.

Pirog, Maureen A., Anne L. Buffardi, Colleen K. Chrisinger, Pradeep Singh, and John Briney. 2009. "Are the Alternatives to Random Assignment Nearly as Good? Statistical Corrections to Nonrandomized Evaluations," *Journal of Policy Analysis and Management*, 28(1): 169-172.

Puma, Michael, Ronna Cook, Stephen Bell, Camilla Heid, Michael Lopez, et al. 2005. *The Head Start Impact Study: First Year Impacts.* Washington, DC: U.S. Department for Health and Human Services. Available at <<http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf>>

Puma, Michael J., Nancy R Burstein, Katie Merrell, and Gary Silverstein. 1990. *Evaluation of the Food Stamp and Employment and Training Program Final Report: Volume 1*. Bethesda, MD: Abt Associates.

Reichardt, Charles S. 2011. "Evaluating Methods for Estimating Program Effects," *American Journal of Evaluation*, 32(2): 246-272.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and non-randomized studies," *Journal of Educational Psychology*, 66: 688-701.

Rubin, Donald B. 2010. "Reflections Stimulated by the Comments of Shadish (2010) and West and Thoemmes (2010)," *Psychological Methods*, 15(1): 38-46.

Schochet, Peter Z., John Burghardt, and Steven Glazerman. 2001. *National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related* Outcomes. Princeton, NJ: Mathematica Policy Research.

Schochet, Peter Z. and John Burghardt. 2007. "Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations," *Evaluation Review*, 31(2): 95-120.

Smith, Jeffrey. 2002. "Evaluating Active Labor Market Policies: Lessons from North America," unpublished manuscript (http://www.glue.umd.edu/~jsmithz).

Tipton, Elizabeth & Larry V. Hedges. 2011. "Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling," presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington, DC.