Journal of MultiDisciplinary Evaluation
Volume 12, Issue 27, 2016

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

On the Feasibility of Extending Social Experiments to Wider Applications

Stephen H. Bell
Abt Associates, Inc.

Laura R. Peck
Abt Associates, Inc.

Background: When deciding how to allocate limited funds for social programs, policymakers and program managers increasingly ask for evidence of effectiveness based on studies that rely on solid methodology, providing credible scientific evidence. The basic claim for the "social experiment"—that the "coin flip" of randomization creates two statistically equivalent groups that do not diverge except through an intervention's effects—makes resulting estimates unbiased. Despite the transparency and conceptual strength of the experimental strategy for revealing the *causal* connection between an intervention and the outcomes of its participants, the wisdom or feasibility of conducting social experiments is often questioned on a variety of grounds.

Purpose: This article defines 15 common concerns about the viability and policy reliability of social experiments, in order to assess how much these issues need constrain the use of the method in providing policy evidence.

Setting: NA

Intervention: NA

Research Design: The research uses the authors' experience designing and conducting dozens of social experiments to examine the basis for and soundness of each concern. It provides examples from the scholarly literature and evaluations in practice of both the problems posed and responses to each issue.

Data Collection and Analysis: NA

Findings: We conclude that none of the 15 concerns precludes substantially extending the use of randomized experiments as a means of evaluating the impacts of government and foundation social policies and programs.

Keywords: *program evaluation; evaluation design; social experiments; internal validity; external validity.*

Introduction

When deciding how to allocate limited taxpayer or donor funds for social programs, policymakers and program managers increasingly ask for evidence of effectiveness based on studies that do not raise quibbles over methodology. They want to assess the extent to which programs have their intended effects 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 policy choices to terminate, continue, or expand social programs should be based on research meeting high standards of evidence. For government and foundation policymakers, strong causal inference showing that a public sector or philanthropic intervention has a favorable impact provides justification for continued funding or expansion. Conversely, unequivocal evidence of ineffectiveness is often needed to justify termination of an existing public program with strong political or bureaucratic constituency but that, with rigorous testing, is found to be producing little or no social benefit. After defining social experiments as a source for credible scientific evidence and discussing briefly their alternatives, this article discusses common concerns about experiments in theory and in practice. It concludes that opportunities exist for extending social experiments to wider applications in appraising the effectiveness of government and foundation social policy initiatives.

What are Social Experiments?

We begin by describing the experimental methodology for measuring social program impacts to ensure it is understood, as a starting point for later addressing potential objections to the methodology that are the focus of the article. 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—for example—receive a new drug while others receive a placebo, an inert dose (or the common standard of care). 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 occur. Because the two groups are *randomly* assigned to their medical treatment experience, the only difference between the two later on is the medication.

Substitute "public policy," "social program," or "intervention" of some sort for "drug" and the same approach applies in testing the effectiveness of public and non-profit efforts to ameliorate social and economic ills. Social experiments deliberately exclude from participation some of the people or organizations an 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 a "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" for what would have happened to participants in the absence of the intervention.

In a successfully implemented experiment, the second condition here 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.

Reliably representing the world without the intervention is crucial to determining whether government and philanthropic social programs make a difference. Experiments that use random assignment—if successfully implemented and effective at meeting the challenges discussed later

in this article—provide a solid counterfactual as represented in the outcomes observed for the control group. This counterfactual allows elimination of the so-called "threats to internal validity"—i.e., plausible rival explanations for why change might occur over time or why differences could arise between participants and nonparticipating comparison groups used as counterfactuals but determined by natural processes rather than random exclusions (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. The experimental design itself is structured to net out all of these influences when subtracting control group outcomes from treatment group outcomes to produce unbiased impact estimates.

This approach runs in contrast to other possible means for approximating a counterfactual in order to determine the contribution of a policy or intervention to changes in social outcomes. Other ways to produce a counterfactual include:

- belief (theology, religious faith, folk wisdom);
- practical experience;
- extrapolation of prior conditions into the future (i.e., presume that no change from preconditions would occur without the policy or intervention);
- measurement of outcomes for "non-treated" cases that occur naturally in the world—cases presumed to be otherwise similar to treated cases; or
- measurement of outcomes for "non-treated" cases deliberately constructed by the researchers through an imposed mechanism or decision rule.

Among the possible mechanisms or decision rules that might be imposed by the researchers are the following:

- a cut-point on a continuous scale above which the policy/intervention is applied (this supports a "regression discontinuity" design and impact analysis);
- "profiling" potential intervention participants on their characteristics as the basis for sorting them into and out of the intervention (this supports a "propensity score matching" design and impact analysis); and
- a random lottery, with treated cases picked by chance from a larger pool—leaving the untreated members of the pool to serve as the counterfactual.

No one would say that the last of these options—using a randomized lottery to produce a social experiment—is everywhere and always the best way to identify counterfactual outcome levels and measure impacts. Indeed, we recognize the longstanding and contentious debate surrounding evaluation methods suited to assessing the causal impacts of policies and programs. We also recognize the recent flurry of "design replication" studies (also called "within-study comparison designs"), in which results from non-experimentally-designed evaluations are tested against experimentally-derived evidence. Although the earliest of social policy replication studies (e.g., Fraker & Maynard, 1987; LaLonde, 1986) reached pessimistic conclusions, as the field has evolved alternative, non-experimental methods have been shown in some cases to replicate the results from their matched experiments in terms of policy implications or, even more reassuringly, in the magnitude of their impact estimates (e.g., Cook, Shadish & Wong, 2008; Cook, Steiner, & Pohl, 2008; Pohl et al., 2009, St. Clair, Cook & Hallberg, 2014; Shadish, 2011). We laud those research efforts but note that until the body of evidence becomes decisive on which circumstances permit specific non-experimental designs to be used with confidence to generate solid causal evidence of policy effects in specific circumstances, experiments will continue to play an important role within the field of evaluation.

Implementing a classically-designed experiment is one method for producing a valid counterfactual, but doing so takes more than a random number generator: enforcing the embargo on program participation among control group members can be especially challenging and often falls short of universal compliance. This, and the failure to obtain follow-up outcome data for all randomly assigned cases (treatment and control), are important challenges to randomized experiments as the paradigm for valid causal inferences. With this as introduction, we now turn to a systematic discussion of some other major concerns about experiments.

Concerns about Experimentation

Despite the transparency and conceptual strength of the experimental design in establishing a causal link between intervention and outcome,

experiments are often questioned on a variety of grounds (e.g., Greenberg & Barnow, 2014). Two of these critiques we have acknowledged as needing concerted attention in implementing and analyzing data from experiments: crossovers of control group cases into the intervention (see Angrist et al., 2006) and incomplete follow-up data (see Puma et al, 2009). Other concerns raised in the research and practice of large-scale impact evaluations apply to all kinds of evaluation designs, and we visit these briefly at the end of the paper. We focus the bulk of the article on 15 concerns that are specific to experiments and, through our analysis and examples, find each one less of an impediment to the use of random assignment research methods for social policy evaluation than is widely believed.

We classify the examined concerns into four categories: ethical, scientific, feasibility, and financial. Across these domains, we argue that none of the concerns examined threatens the reliability or viability of experimental techniques for measuring the impacts of social interventions. As a result, we conclude that government agencies and foundation funders have the opportunity to use experimental methods to obtain transparent and compelling answers to important social policy impact questions in more instances than they may recognize. While other impact analysis methods may be appropriate in certain circumstances, not until two things happen would we recommend wide use of non-experimental impact analysis approaches: evaluators gain a better understanding of intervention participation selection processes in the complex systems in which social policies operate and a larger body of design replication studies inform the circumstances in which selected non-experimental designs give accurate findings.

Recognizing that experiments 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. Policymakers and program administrators can then confidently use information about the consequences of the variable factor—the social intervention being tested—to decide on the intervention's future use.

That said, experiments face a variety of challenges, leading us to three areas of inquiry: What are some major concerns or criticisms raised about social experiments? To what extent are those concerns valid? To what extent are these criticisms surmountable? The viewpoints expressed in our exploration of these questions—including the identification of the specific concerns about social experiments brought up for discussion—reflect decades of collective experience designing and analyzing randomized and non-randomized comparison group 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). While some of the concerns are well-cited in existing scholarship, some of them are not, instead being identified from practical experience. As such, this is the first time some of these criticisms appear in the literature.

The discussion and analysis offered are intended to spur dialogue among policy evaluation researchers and funders and push the field forward (Bell, 2003). The primary audience is government and foundation funders who decide what type of evaluations to undertake. It is important that knowledge from the field, reflective of the latest, most complete experience of evaluators, reach those decision-makers. It is also important that practitioners of impact evaluation understand what can be accomplished with experimental evaluation designs so they can support funders in carrying out experiments when called for.

To serve both of the funder and practitioner audiences, we now review 15 concerns about randomized experiments and why we believe each can be overcome.

The Ethical Concern

Social experiments must answer the most fundamental challenge to their legitimacy—the contention that exclusion of some eligible and deserving individuals or organizations from a program's services or a policy's provisions for the sake of research is unethical. We address this criticism from a variety of perspectives in this section.

Concern #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; Blustein, 2005). Are the

individuals who "lose the government lottery" and enter the control group disadvantaged unfairly or unethically? Some government programs are entitlements; denying access to them for the sake of research would be not only unethical but also illegal. Of course no evaluation should ever propose illegal treatment of potential research subjects. While the issue of ethics 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 and Shroder (2004) catalogue over 200 of them, plus the *Randomized Social Experiments eJournal's* early 2015 count of at least 500 experiments since 2004. So at some level social experiments are ethically acceptable for the American polity; likewise, experimental evaluation designs are in wide use internationally, including in developing nations. Why might this be the case? The main ethical concern is that randomizing people to a control group denies them access to opportunities that they would otherwise have had and that could potentially benefit them. Three responses to this concern might be considered.

First, if—due to funding or administrative capacity constraints—a program has to limit the total number of people or organizations served relative to the number that seek services, *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 services that cannot be provided to all (e.g., Bickman & Reich, 2009; Orr, 1999). Surely it is more fair than allowing program staff to choose their favorite applicants based on personality, personal connections, or perceptions of who would benefit most from participating—or than serving those who happen to apply when service funding is flush rather than when it is scarce. It can be argued, then, that oversubscription to a program makes random assignment ethical.[1]

Second, if a program's effectiveness has yet to be determined, *being turned away from participation as part of a control group should not be presumed to be more detrimental 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, *because the answer is not known*, participating in it at best constitutes a neutral situation and may be disadvantageous, at least for the time it wastes. One example of this is the U.S. Job Training Partnership Act (JTPA) program of the 1990s. A large-scale randomized impact evaluation found this program to cause unemployed youths to wait longer to go to work than their counterparts in the experimental control group (Orr et al., 1996), possibly because they expected the 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, science should presume *no* impacts until proven otherwise. Peter Rossi's "Iron Law of Evaluation" is that "the expected value of any net impact assessment of any large scale social program is zero" (1987, 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 not start from the position that placement in the control group will put individuals or organizations behind where they would have been had they been assigned to receive program services. True, new ways of providing social assistance only get legislated, or put forward as demonstration tests, when someone believes they will be beneficial. Whether this is in fact the case is unknown and is the rationale for considering a rigorous experimental test of the intervention. It is unknown whether being in a randomized control group would cause *harm*, in the same way that it is unknown whether being in the treatment group would cause benefits. This uncertainty is the whole reason for conducting an experiment in the first place.[2] Of course, one should not exclude a control

[1] The case of researchers *inducing* more individuals or organizations to apply for services in order to create oversubscription is less clear. In that situation, some individuals or organizations that would have remained disinterested in or unaware of the program's services decide they would like them. Those assigned to the treatment group cannot be harmed by stirring up interest, since the resulting "demand" for services is met. The corresponding control group members may be temporarily harmed by having their hopes—and application costs—raised only to be denied services in the end. In this case different bases for arguing the ethical acceptability of random assignment must be found, such as those noted below.

[2] Notwithstanding this fact, a small group of individuals assigned to the control group in the National Job Corps Study (Schochet, et al., 2001) sued for having been harmed by exclusion from the treatment group and won

group from a beneficial intervention for research purposes a second time once proof of a benefit is in hand, nor keep them in the "untreated" condition should an ongoing experiment show benefits of a potentially life-altering magnitude.

A similar point applies to medical trials: early checks for effects on mortality and other extremely consequential outcomes are thought essential for ethical reasons, as is the imperative to stop withholding life-saving treatments from the control group once early impact findings indicate that lives can be saved or major negative health consequences averted. Randomized experimental tests of social policies with the potential for major benefits of this magnitude (if there are any) should always include collection, analysis, and reporting of early outcome data to ensure that control group members are not deprived of vital benefits any longer than is necessary to discover that those benefits exist. Believing in or hoping for any kind of benefit is not the same thing as proof. If harm as well as benefit could arise from an unproven intervention, 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 course to follow, as Blustein (2005) elaborates. Society, which benefits from accurate information about program and policy effectiveness, may be justified in allowing some individuals or organizations to be disadvantaged in order to gather that information and thereby achieve much wider benefits. Society regularly disadvantages individuals based on government policy decisions undertaken for non-research reasons, such as when free trade agreements cost some workers their jobs while creating jobs for others in export industries or when the opening of high-occupancy vehicle (HOV) lanes disadvantage solo commuters to the benefit of carpool groups. Moreover, unlike changes in other societal rules, the potential losses from assignment to a control group for research reasons are temporary—until the end of the study period—while the benefits of good information to society will be long-term. As in medical research, if major benefits of a treatment are quickly proven, then a social policy experiment should enroll control group members in the intervention without waiting for the study to run its originally planned course.

No researcher can appropriately weigh the balance of these considerations on society's behalf when considering whether the broad use of randomized impact studies as a way to improve policy is justified. That is a decision that must fall to public officials. Still, on the basis of the above considerations, the question of the inherent fairness or lack of fairness of random assignment remains an open question not a firm impediment to the use of the method, with arguments to be made on both sides of the issue.

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 and to evaluations of pilot tests of new interventions. Some of these concerns have explicitly arisen in the scholarly literature, which we cite, while others have not, meaning they are undocumented issues that have emerged from the authors' own examination of possible scientific limitations of randomized social experiments.

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

The first scientific criticism of randomized experiments that we wish to consider 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 ITT/TOT distinction arises whenever less than 100 percent of a randomly assigned treatment group participates in the assigned intervention—i.e., when the "treated" group is different (smaller) than the full experimental sample assigned to treatment. Less than 100 percent participation of treatment group members is common in experimental evaluations: individuals cannot be compelled to take part in an intervention such as subsidized housing simply because they applied for it and were randomly selected to be offered admission. Applicants for government social programs regularly change their minds about participation before the admission decision comes, a result inevitable as well then for

their case, providing some voice in an otherwise relatively silent debate about the ethics of random assignment in social policy evaluation (Blustein, 2005).

some number of applicants chosen by random assignment.

On the one hand, the ITT estimate of the average effect of being assigned to receive treatment might be considered the most policy-relevant measure of a program's impact.[3] It reflects the target population's overall response to the intervention—the *combined* consequences of (i) decisions to participate and (ii) impacts once participating—for the target group *as a whole*. Arguably, this is what policy-makers need to know in some contexts. The ITT estimate essentially averages impacts across those who took up the offer of treatment and those who did not; it represents what is likely to happen on average in the target population as a whole from future offers of the same intervention.

On the other hand, the TOT estimate tells policymakers the average impact of participating in the intervention for those members of the target population who choose to participate after receiving the offer. Impact information on this subset provides the best comparison to average spending for the intervention—cost per funded "slot"—for funders and is presumably what prospective participants want to know: how much they can expect to gain if they do participate. So TOT estimates are of interest to both policymakers and a broader population. Therefore one might wish that social experiments were capable of producing both types of impact estimates reliably, ITT and TOT.

One can calculate the ITT impact measure without bias in an experiment as the difference between the average outcome for all treatment group members (all of whom receive the intervention offer, though some do not take it up) and the average outcome for all control group members. Fortunately, the TOT estimate is also readily obtainable from the experimental data without bias, subject to an assumption. Applying what is known in the literature as the "no-show adjustment" converts any ITT estimate into the corresponding TOT estimate. Introduced by Bloom (1984), the no-show adjustment requires that *the intervention has no effect on members of the treatment group who do not participate*—for example, the students 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 observers—evaluators and policymakers—when voluntary participation interventions are examined and reliable data on participation can be obtained.

Based on this assumption, the initial measure of impact—the intervention's average impact across all treatment group members in the ITT estimate—includes both potentially positive (or negative) effects on participants and zero effects on non-participants (the "no-shows"). Such an estimate can be rescaled to remove the diluting effect of non-participation, providing a measure of the average impact *on just those who do participate*. No assumptions regarding the similarity of participants and non-participants is needed, nor is the ability to adjust statistically for differences between the two groups; the participants and "no-shows" can be as different as night and day and the TOT estimate based on Bloom's method remains unbiased as long as the intervention indeed had no effect on the "no-shows."

Concern #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, a social experiments' counterfactual is usually described as the "*status quo*" or as "business as usual." Random assignment in social policy evaluations does not control whether individuals access similar alternative services; as a result, some control group members inevitably do so. This means that the control group is not a "no services" placebo in most social policy evaluations. 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 and access to one or the other is randomized for research purposes. Increasingly, medical and policy evaluations alike are avoiding no-services control groups and instead testing the relative effects of contrasting treatments.

The circumstance of evaluating a program relative to everything else that exists in the

[3] This average effect can also be quite small and hence difficult to detect statistically, when only a small percentage of the treatment group receives the assigned intervention—a circumstance in which most individuals experience zero impact. We classify this as a feasibility issue rather than a scientific validity issue and consider it as Concern #10 below.

community gives randomized impact studies the same character as the real-world programs they seek to evaluate. As such, this is a strength, rather than a weakness, of the experimental approach as long as the nature of usual services if fully documented. 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. When precisely this event occurs for some of the children assigned to the control group in the Head Start Impact Study (Puma et al., 2005) it is a good thing, from the standpoint of understanding Head Start's contribution to *improving* on what services are otherwise available in the community. It is precisely the choice between the experiment's two scenarios—a set of children enrolled in Head Start (the treatment group), or a subset of the same children served by state pre-kindergarten programs (represented by the control group)—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 that program's impact. Knowing how a given intervention compares to no intervention at all does not help social decision-making in a fragmented federalist system with many intervention sponsors and selective consumer participation among available programs. 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 is out there is enough, then the money spent on DHHS programs could be cut back without consequence. The same is true for other programs sponsored across the range of social policy areas.

Concern #4: Counterfactual experiences in the control group are distorted by easier access to alternative services (i.e., services from sources other than the program being evaluated).

This concern—which we have noted in previous unpublished work of our own but have not seen previously in the literature—arises from the possibility of "queuing effects" among control group members. The *existence* of the intervention under study *may shorten the queue for control group members seeking to access alternative services*, compared to the competition for those services they would have faced if the studied program did not exist. Because treatment group members (and other, non-research individuals) participate in the program under study, they do not contend with control group members for access to outside services, as would happen in the world the control group is supposed to represent—a world in which the studied program does not exist. With competition for similar services from other sources less fierce than it would be in a true counterfactual world, control group members get too much help from other sources.

To see more clearly 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 Innovation and Opportunity Act (WIOA) program, was completely eliminated. In this scenario, the total supply of employment and training services in the nation would fall precipitously. Everyone seeking employment and training assistance, including those ordinarily served by WIOA, would have to scramble for the available services "slots." But because in an actual social experiment WIOA *would not* go away, control group members from such a study would not have to scramble in this way with such a larger group of other training-seekers. In turn, more of them would find "slots," potentially altering treatment/control differences in outcomes and producing an undersestimate of WIOA's impact.

If Congress were deciding between continuing versus eliminating WIOA, one would want to run a social experiment in which (1) treatment group members are given access to WIOA, and (2) control group members compete for access to training services from non-WIOA sources, but do so in a "market" in which treatment group members are also vying for those alternative training slots. Unfortunately, the second condition cannot be met: the treatment group cannot simultaneously participate in the WIOA program that still exists and jostle with the control group for access to the limited number of alternative service sources, sometimes squeezing them out of those slots.

This critique hinges on two unstated assumptions: that evaluation results will guide a decision to either keep WIOA at its current scale or eliminate it altogether, and that other programs providing similar services to the same customer group *would not expand their scale were WIOA eliminated*. If Congress were choosing between full-scale WIOA and no WIOA and did not expect the "hole" WIOA would leave if discontinued to be

filled by other employment and training programs, we would indeed want control group members to have to compete with treatment group members for other, non-WIOA training slots to achieve the appropriate counterfactual for the policy choice at hand. But if Congress expects other funders to expand services in response to the "shortages" created by WIOA's disappearance, then expanded service availability for the control group—up to a point—represents the correct counterfactual.

A similar point holds when deciding whether to expand or contract WIOA funding *at the margin*. Such a change would affect only a small share of all those seeking WIOA-type services. In this circumstance, what happens to control group members will represent the desired option and outcomes well—they really would not have to compete with many more workers for training slots when the WIOA program size changes only fractionally. Therefore, the contrast produced by the treatment-control comparison in the actual experiment would reasonably trace the consequences of the considered policy decision.

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 at least 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 to those services. Neither is perfect, but in principle the perfect version of control group experiences is unknowable until policies actually change—either marginally or dramatically—and other agencies react—either a little or a lot. Absent that information, a cautious approach appropriate to marginal changes and the approach that social experiments naturally produce in our judgment provides the safer basis for policy assessment.

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

Another similar, 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: either it serves fewer people, operating below capacity (or, if below capacity anyway, operating even further below capacity than usual) or it serves additional people who ordinarily would not be served due to capacity constraints. 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 brings in others who normally would not participate.

The question is whether either of these results matters to the size of the program impacts measured compared to the quantity one wants to measure via random assignment? Likely both results do matter, though perhaps not to a very great extent. A program with added vacancies created by random assignment may deliver services differently for the customers it does serve. If budgets remain unchanged, then the typical participant in a less fully-subscribed program presumably receives more services and hence may experience a larger impact. Or smaller numbers of participants may change the dynamics of 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 as part of unnaturally small groups.

Alternatively, program scale and operations could remain unchanged if added people are served who 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. They may even be people who would not have applied to participate, if the program has to expand its recruiting to supply enough applicants for a control group and remain fully subscribed.[4] 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

[4] An added recruitment effort itself might change how the program does other things, including the nature and effectiveness of the services it delivers to treatment group members. This would be another distorting effect of running an experiment.

from the program's services (which might or might not be the same people) do. On one factor or another, entrants differ from non-admitted applicants and the resulting differences could lead to larger or smaller program impacts. The intervention (treatment) and counterfactual (control) samples continue to match one another, through random assignment, *but collectively they represent the wrong set of people, a somewhat different and larger set than would ordinarily be served.*

Fortunately, Olsen, Bell and Nichols (2016) propose a way to identify which individuals would ordinarily have been served so that impact results can be produced for just that subset. Under their approach, local program operators are given the opportunity to identify applicants they would enroll in a normal year (i.e., a year without a randomized admissions lottery). Their incentive to do so reliably results from another design twist: increasing the probability of being assigned to the treatment group for "normal year" enrollees, something program managers presumably would desire (since these are the applicants they would choose to admit first in a normal year). This set of participants and their control group counterparts can be analyzed as a subgroup defined by pre-random assignment information (normal-year enrollee/normal year non-enrollee status) and the study can obtain impact estimates for the normally-served population.[5]

No good data exist on how much this would 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 people it normally would serve. The National Job Corps Study provides an excellent example of steering clear of distortions to the treatment group by this means—making very few control group exclusions in any one program site—while still achieving a large overall control group sample through inclusion of many sites (Schochet et al., 2001). This model should be emulated elsewhere.

Concern #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.

The final scientific objection we consider is that experiments eliminate selection bias in measuring the effect of services provided from *the point of random assignment forward* but do not provide equally strong information on the consequences of services provided at other stages of the intake and in-program service delivery processes. Experiments instead leave evaluators with nothing to turn to answer other policy questions besides less reliable non-experimental comparisons. For example, consider the relative impact of varying sequences or "dosages" of services determined after randomization (and never observed for the control group). One might like to examine how much difference these program elements make to program success; but, without randomizing, these in-program participation patterns are endogenous, and the analysis of their effects are likely to suffer from selection bias. 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 treatment and control group members.

This discussion emphasizes the importance of choosing wisely when deciding where to position random assignment within a program's intake flow. Does this issue point out a weakness of experimental designs compared to other impact analysis strategies that yield estimates of the effects of multiple program features? No: researchers using experimental data have all the same options available for non-experimental analyses of impacts of program components as non-experimental studies (think of simply "setting aside the control group data"). Moreover, techniques exist that capitalize on the experimental design when defining and analyzing subgroups defined by post-random assignment events, choices or program features (e.g., Peck, 2013, 2015).[6]

[5] A bonus of the Olsen, Bell and Nichols (2016) method arises from the normal year non-enrollee subgroup included in the experiment. Analyzed separately, this population yields an estimate of the effect of program expansion on the marginal enrollees, something often of great importance to funding expansion decisions.

[6] These analytic techniques (see for example Peck, 2003, 2005, 2007, 2013, 2015; Bell & Peck, 2013; Harvill, Peck, & Bell, 2013) reduce the seriousness of concern #6 by providing opportunities to explore questions of

In sum, at the place where randomization is inserted in a social experiment the research gets stronger. Everywhere else it remains the same. Eliminating one selection bias problem—the one that distorts the answer to the most important policy question to be addressed by a study—is clearly a virtue of experiments compared to eliminating none. Moreover, random assignment at more than one point in the intake and service delivery flow will give strong experimental answers to more than one policy question within a single randomized study.

Feasibility Concerns

Beginning with the alleged ethical and scientific failings of randomized experiments, we have argued that these concerns are not the terminus for experiments but instead issues on which to focus in building stronger impact evaluations. The next step would seem straightforward—"Just do them"—but it is not. Feasibility issues also demand consideration. Can a scientifically valid, ethically acceptable research approach actually be pulled off reliably under real-world conditions? Our conclusion in this realm is that eight feasibility concerns can be overcome by devoting sufficient resources to sizing and conducting an experimental evaluation, if the policy question to be addressed is of sufficient importance. Naturally, the policy questions of insufficient importance to address one or more feasibility issue through more spending should not be studied with experimental methods (see discussion of funding tradeoffs at Concern #14 below).

Concern #7: Saturation interventions that affect entire local communities cannot be randomly assigned.

Evaluations of systems change and other community-wide "saturation" interventions are almost universally evaluated using non-experimental methods (e.g., Connell, Kubish, Schorr, & Weiss, 1998; Fulbright-Anderson, Kubish, & Connell, 2002; Nichols, 2013). The one known exception (Bloom et al., 2005) randomized a small number of comparatively small communities, 15 public housing developments in six cities. While it has become common to randomize clusters of study subjects instead of individual units—for example, clusters of students and/or classrooms (i.e., full schools) in educational evaluations (Jacob, Zhu & Bloom, 2009) or clusters of patients or physicians (i.e., clinics) in health policy evaluations (Meurer & Lewis, 2015)—these are not communities in a geographic sense.

We consider evaluations of community-wide interventions to be prime candidates for application of the experimental method, if the policy questions to be addressed are sufficiently important to justify the resources required. The U.S. is a very large nation, with tens of thousands of local communities or neighborhoods that could be randomly assigned into or out of a particular community-level policy or intervention. Likewise, the world is large, with many places, both within and across countries, eligible to be the subject of place-based policy evaluation. There is no feasibility constraint to randomizing across many places, only a willingness constraint—one that society can overcome if sufficient importance attaches to the policy question posed.

Community-wide saturation interventions make data collection more difficult and expensive, and any impacts that do occur harder to find because they tend to be diffused across many people in the community. These drawbacks afflict *any* impact evaluation research on saturation interventions, not just experiments. Therefore, the simple fact that an intervention involves community saturation is not a sufficient argument in our judgment to dismiss using an experimental design to evaluate its impacts.

Concern #8: Programs that struggle to meet enrollment targets cannot provide the sample sizes needed for a randomized experiment, once the total number of eligible applicants required goes up with the addition of a control group.

Concern about meeting the sample size requirements of experimentally-designed evaluations understandably arises in situations where programs are already struggling to meet their enrollment targets. How can such programs allocate a large number of eligible applicants to an unserved control group and still fill up all of their funded service "slots"? One possibility is that programs in this situation are not sufficiently in demand in their communities to warrant their

dosage or varying treatment paths within the experimental design. For a recent applied example, see Peck and Bell's (2014) analysis of the role of quality in children's Head Start experiences, or Moulton, Peck & Dillman's (2014) analysis of the role of neighborhood quality in the Moving to Opportunity (MTO) demonstration.

current level of funding and therefore should have their funding reduced at the margin, at which point an experimental evaluation is no longer infeasible. However, it may be important to have information on program effectiveness before funding reductions are considered. In that case, only a small number of control group cases needs to be sampled in any locality compared to the size of the overall program in that community. As long as enough localities can be included in the study, this will still provide adequate sample sizes for the evaluation. The Head Start Impact Study took this approach, requiring only 11 control group members per Head Start center to be included in the evaluation—centers that typically served 100 or more children—while including over 300 centers in the study (Puma et al., 2005).

In some instances, additional technical assistance resources may be needed to increase the flow of applicants sufficiently to accommodate a modest-sized control group without leaving funded slots unused. However, this raises the possibility of a compositional shift in the population served discussed in conjunction with Concern #5 above.

Concern #9: Randomized experiments are 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' decisions in areas typically considered under long time horizons where long-term planning guides behavior, such as decisions to return to work in the face of permanent disability or to invest in a new production plant capacity. We would not expect policies intended to affect long-run behavior to reveal their full impact in an experimental setting unless the treatment group members in the 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 here is that these are unrealistic conditions to impose when testing a new policy intervention—i.e., that a demonstration project cannot create a credible sense of permanency for either group: control group members may come to expect their "turn" to get the new intervention 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 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." Second, policy conditions for treatment and control group members in an evaluation do not have to be unnaturally abbreviated simply 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 tested policy is sufficiently important to its success.

One example is the provision of alternative disability benefits to individuals whose medical conditions are 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 lifetime 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 during the interval in which the impact assessment takes place, control group members *were not expecting* to ever receive the new services, and treatment group members were confident of the *permanency* of their policy provisions.

Concern #10: Randomized experiments are not appropriate for interventions with low participation following randomization, because average effects for the treatment group as a whole will be too small to be detected.

The discussion here extends consideration of the ramifications of less than 100 percent participation begun at Concern #2 above. There, the issue was whether incomplete participation results in experimental impact analysis answering the wrong policy question. Here, the policy question is presumed to be on target but concern arises about a study's ability to detect non-zero

impacts of policy importance as statistically significant. The simple fact is that small average effects can be detected in sufficiently large samples. The cost of the larger samples is justified if the anticipated impact (i) would be of great policy importance even if small on average, or (ii) becomes large enough to hold great policy importance once translated into the effect of the intervention on the average participant.

Concern #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, to determine *which features of the program account for its effectiveness*. This second category of information allows funders to make interventions more cost-effective by increasing the effective components and/or eliminating components that do not add to impacts. The up/down nature of experimental findings concerning the entire package of intervention components under study is thought to severely limit the usefulness of social experiments as a way to discover how a program can be made *more* effective or less costly through changes in the intervention elements.

One response to this criticism of evaluations that use random assignment is obvious: randomize more things, including the components of intervention packages or different variants of an overall intervention approach. Recent examples such as the U.S. Department of Health and Human Services' evaluation of a health sector career pathways-based training program (the Health Profession Opportunity Grant [HPOG] program), includes randomization to two treatment arms and a control group, where the treatment arms include the basic HPOG intervention and an enhanced version of the intervention. This allows determining the extent to which the added features of the enhancement are important 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 and Robins, 1986). While this approach may sacrifice statistical precision if not sufficiently powered (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 randomized experiments. Recent scholarship has highlighted many design opportunities that can help answer these more nuanced questions (Peck, 2016a).

Moreover, *multi-stage* random assignment can be used to answer questions about the effects of different treatment experiences without sacrificing statistical precision (Bell & Peck, 2016). For example, suppose a government agency wanted to know if a work incentive would increase employment among people receiving public income support benefits, and whether the impact of such incentives on employment success and self-sufficiency would be increased by providing those induced by the incentives to go to work with additional job 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 new work incentive, another treatment group receiving the new work incentive plus added job supports, and a control group—thereby cutting the size of each group by one-third relative to two-arm random assignment—an innovative design would randomize at two different points in time: two-way randomization of the full sample to determine which individuals receive the new work incentive, followed by—for individuals in the incentivized group who obtain jobs—separate two-way randomization to the provision or non-provision of the added job supports. This design not only increases the 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-job-supports policy), it concentrates the examination of the impact of job supports on just individuals who, if given the opportunity, actually use them—making statistically detectable impacts more likely.

We further elaborate on these points about creativity and flexibility in randomized experimental designs in Bell and Peck (2016). 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 Concern #6 above apply here as well. These methods capitalize on the strengths of a randomized design while adding minimally intrusive assumptions (and sensitivity tests thereof). By these means, one can learn which components matter, without further randomization among various treatment regimes. The criticism that social experiments cannot "look inside the black box" to determine which elements of an intervention generate its success ignores design and analysis alternatives capable of revealing which components of multi-faceted interventions lead to their success. For further elaboration of "black box" opening designs and analyses, see the recent special section on this topic, published in the *American Journal of Evaluation* volume 36, issue 4 (2015) and large portions of issue 152 of *New Directions in Evaluation* on Social Experiments in Practice (Peck, 2016b).

Concern #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 inevitably spill over to individuals not directly engaged in the program or services offered. This happens, for example, if job training equips workers to take jobs other workers would otherwise have held, resulting in earnings losses for workers not included in the research. Economists call this kind of situation a "general equilibrium effect" because it ripples through the social or economic system creating impacts in secondary locations not covered by the direct research sample. Smith (2002) provides several good examples of these general equilibrium effects of labor market interventions and explores the evaluation challenges they create.

Not only randomized experiments but all research based exclusively on data for individuals participating in an intervention and a confined sample of non-participants, such as an experimental control group, faces this issue—spillover outside the study group. Randomization does not make these spillover effects *more* difficult to measure. The wider "ripple effects" of social policy initiatives confronted by general equilibrium analyses are always difficult and potentially expensive to measure, no more so for having measured the direct effects of those policies experimentally.

Concern #13: Experiments have limited generalizability since usually they are not based on a statistically representative set of sites.

The appeal of randomized trials centers on their "internal validity"—i.e., their ability to provide unbiased information concerning impacts on the people or organizations directly studied and hence *internal* to the study sample. Policymakers also care about the "external validity" of evaluations—i.e., whether study findings are accurate for some larger *external* universe of interest, such as all homeless families in America or all minority-owned small businesses. Researchers have argued that social experiments that are not specifically designed to provide external validity may give bad policy guidance for the pertinent population even when they provide very accurate evidence on impacts within the study sample (Olsen et al., 2013).

At least a half-dozen experimental impact evaluations of ongoing social programs have achieved both internal and external validity, the latter by being conducted in geographically-based probability samples of the nation (without substantial attrition of local programs from the research) that formally represent all Americans. The Food Stamp Employment and Training Evaluation (Puma et al., 1990) is one example, the Head Start Impact Study (Puma et al., 2005) another. But what about social experiments that do not have external validity—are they still worth doing and to be preferred to non-experimental impact evaluations with external validity but that—due to selection bias—lack internal validity? Scholars have debated the internal versus external validity tradeoff since these two terms were invented (e.g., Bracht & Glass, 1986; Jimenez-Buedo & Miller, 2010), with a general consensus preferencing internal to external validity (Reichardt, 2011)—and hence favoring experiments. But this does not put external validity out of reach of social experiments, as emphasized in recent work (e.g., Bell & Stuart, 2016; Olsen & Orr, 2016; Tipton et al., 2014; Tipton & Peck, 2016).

Concern #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). This is not a point unique to experiments: if policymakers are interested in long-term outcomes and impacts—or in interventions that themselves take a long time to administer—then any prospective evaluation design will require the time needed to cover the policy-relevant follow-up period. As a challenge to experiments in particular there is a potential response. Government agencies have the option of establishing a system of regular and temporally-overlapping experimental evaluations of ongoing programs so that new information is always emerging from experimental data.

In addition, evaluations of shorter-term impacts 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 studied 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

Concern #15: Experiments are too expensive.

The eight feasibility concerns raised in the previous section 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 therefore indirectly to taxpayers or foundation donors) have often been put forth as an important obstacle to use of experimental designs. It is beyond the scope of this article to investigate the costs of alternative impact evaluation designs in any detail. Suffice it to say that budgetary constraints on funding agencies—be they government, foundation or nonprofit—are not valid reasons 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 never inexpensive unless the data come from administrative records routinely compiled for program management reasons such as state Unemployment Insurance wage records or homeless shelter occupancy rosters. Whether the critical data come from existing research or new surveys, this facet of research cost is invariant to *how* program participants and non-participants are selected.

Data for non-experimental impact analyses can also be taken from large surveys of households and workers collected for purposes apart from impact evaluations of specific programs, such as the Current Population Survey and the Survey of Program Dynamics. Were such evaluations reliable sources of policy guidance, the fact that the social costs of data collection have already been paid would allow individual federal agencies to use the information at low marginal cost. However, national surveys combined with non-experimental analytic approaches was the first strategy for measuring the impacts of social programs non-experimentally discredited by careful methodological research in the 1980s (Barnow, 1987; LaLonde, 1986).

Small local reforms can also be examined experimentally using low-cost data. Researchers have argued that such reforms can provide an ideal testing ground for incremental changes useful for refining policies and programs and subsequently improving their performance. For example, Peck and Scott (2005) document one state's efforts to change its public assistance intake process by this means. Not a major reform, the initiative used an experimental design to ascertain the extent to which welfare recipients were 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 and provided data on treatment and control group cases' characteristics and outcomes, and university researchers analyzed the data to determine short-term impacts. The costs of this pilot test were minor 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 large national policy reforms. The Coalition for Evidence-Based Policy (2007), its then president (Baron, 2012), and more recent efforts by the Laura and John Arnold Foundation (2015) all reinforce the point that experimental evaluations need not be costly.

On a larger scale, society must consider the "opportunity costs" of *failing* to do experiments—i.e., the money spent on ineffective programs that continue to be funded (and continue to offer false hope) because unbiased information on their inadequate impacts has not been produced. From this perspective, experiments may be the comparatively *low cost* option compared to other impact evaluation approaches once an appropriately broad social viewpoint is adopted (e.g., Burtless & Orr, 1986; Orr, 1999).

Importantly, prominent researchers known for their contributions to non-experimental impact evaluation methods have 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 and Shroder (2004), in the *Digest of Social Experiments*, provide an important closing perspective on the cost issue: "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" (p.13). It would be difficult for today's national government to back away from the practice of rigorous, experimental impact evaluation on the grounds of insufficient funds, when reliable policy guidance is known to depend on the use of randomized experimental designs.

Other Concerns

In addition to the 15 concerns about social experiments just discussed, other scientific and practical limitations face all large-scale policy impact evaluations of *any* design, experimental or non-experimental. These issues 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 variation in apparent impacts across subgroups and locales.

Another criticism of social experiments concerns incorrect analysis of nested or hierarchical data, if the analytic method is not aligned to the level of the hierarchy at which randomization occurs. Research on education reform provides one example: entire schools randomized in and out of the treatment but impact analysis that produces 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 because 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), we expect this issue to disappear.

Discussion and Conclusions

The purposes of this article are to identify some perceived limitations of experimental designs when researching social program impacts—whether those limitations arise from past publications or practice—and to explore the extent to which these concerns preclude the use of experiments in a range of social policy evaluation contexts. In terms of the primary ethical criticism, we argue that an equal-opportunity lottery is the fairest way to ration access to supply-limited social services. Running such a lottery for research purposes 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 to help disadvantaged citizens going forward. The several scientific objections that have been raised about social experiments also appear to us unfounded or minor relative to the strength of the randomized evaluation approach and its potential to inform policy decisions about program effectiveness. The many feasibility obstacles to experiments on closer inspection appear to be only "speed bumps" that do not block the road toward greater use of social experimentation and can be overcome with adequate funding. Finally, we conclude that the reluctance to bear the perceived higher cost of social experiments compared to

impact evaluations with non-experimental designs 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 this discussion of potential pitfalls in social experimentation is useful to those in the government and foundation sectors planning future evaluation activity. We also hope that some of the factors thought to be obstacles to using social experiments receive greater debate in the literature and other forums beyond this article, and perhaps as a result come to be seen more clearly. A reassessment of the strengths and limitations of randomized experiments seems particularly appropriate at this juncture in light of recent advances in experimental evaluation design and analysis that have strengthened the ability of experiments to address historical concerns in ways cited in this article.

To recap our thesis: Are experiments sufficiently robust to serve as the customary standard of practice in social program impact evaluation? Operationally and scientifically, we believe they are, particularly if the political will and funding commitment 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—a topic that deserves more careful inspection. In the meantime, the argument of this article is that issues of ethics, scientific integrity, and practical 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, false alarms.

Acknowledgements

The authors are grateful for support from the Abt Associates Senior Fellows program. We also acknowledge useful input from Abt Associates' Journal Author Support Group, and conference discussants and attendees at the International Conference on Field Experiments in Policy Evaluation (Nuremburg, Germany) and the Southeast Economics Association (Tampa, FL USA), especially Burt Barnow. Any errors in substance or logic remain our own.

References

Angrist, J. D. Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455. DOI: 10.2307/2291629

Barnow, B. (1987). The impact of CETA programs on earnings: A review of the literature. *The Journal of Human Resources*, *22,* 157-193.

Baron, J. 1 (2012). *Rigorous program evaluations on a budget: how low-cost randomized controlled trials are possible in many areas of social policy*. Washington, DC: Coalition for Evidence-Based Policy.

Bell, S. H. (2003). *Review of alternative methodologies for employment and training research*. Washington, DC: U.S. Department of Labor, Employment and Training Administration Occasional Paper 2003-11.

Bell, S. H.., & Peck, L. R. (2016). On the "How" of Social Experiments: Experimental Designs for Getting Inside the Black Box. *New Directions for Evaluation, 152*.

Bell, S. H., & Peck, L. R. (2013). Using symmetric predication of endogenous subgroups for causal inferences about program effects under robust assumptions: Part two of a method note in three parts. *American Journal of Evaluation*, *34,* 413-426. doi: 10.1177/1098214013489338

Bell, S. H., & Stuart, E. A. (2016). On the "Where" of Social Experiments: The Nature and Extent of the Generalizability Problem. *New Directions for Evaluation. 152*.

Besharov, D. (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,* 200-222.

Blustein, J. (2005). Toward more public discussion of the ethics of federal social program evaluation. *Journal of Policy Analysis and Management*, *24*, 824-846.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*, 225-246.

Bloom, H. S., Riccio, J. A., Verma, N, & Walter, J. (2005) *Promoting work in public housing: the effectiveness of Jobs-Plus, final report*. New York, NY: MDRC. Available at <<http://www.mdrc.org/sites/default/files/full_485.pdf >>

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

Boruch, R. F., Victor, T., & Cecil, J. S. (2000). Resolving ethical and legal problems in randomized experiments. *Crime & Delinquency*, *46*, 330-353. doi: 10.1177/0011128700046003005

Burtless, G., & Orr, L. L. (1986). Are classical experiments needed for manpower policy? *The Journal of Human Resources*, 21(4): 606-639.

Bracht, G. H., & Glass, G. V. (1968). The External Validity of Experiments. *American Educational Research Journal*, *5*, 437.

Coalition for Evidence Based Policy. (2007). When is it possible to conduct a randomized controlled trial in education at reduced cost, using existing data sources? A brief overview. Available at <<http://www.evidencebasedpolicy.org/docs/PublicationReducedCostRCTsUsingAdmin07.pdf>>

Connell, J. P., Kubisch, A. C., Schorr, L. B., & Weiss, C. H. (eds.). (1998). *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*. Washington, DC: Aspen Institute.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research, in Mosteller, F., & Boruch, R. F. (eds.) *Evidence Matters: Randomized Trials in Education Research*, Chapter 6, pp. 150-178.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724-750.

Cook, T. D., Steiner, P. M., & Pohl, S. (2008). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research*, *44*, 828-847.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources, 22*, 194-227.

Fulbright-Anderson, K., Kubisch, A. C., & Connell, J. P (eds.). (2002). *New approaches to evaluating community initiatives, Vol. 2: Theory, measurement, and analysis*. Washington, DC: Aspen Institute.

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

Gueron, J. M. (2002). The politics of random assignment: Implementing studies affecting policy, in Mosteller, F., & Boruch, R.F. (eds.), *Evidence Matters: Randomized Trials in Education Research*, Chapter 2, pp. 15-49.

Greenberg, D., & Barnow, B. (2014). Flaws in evaluations of social programs: Illustrations from randomized controlled trials. *Evaluation Review*, *38*, 359-387.

Greenberg, D. H., & Robins, P. K. (1986). The changing role of social experiments in policy analysis. *Journal of Policy Analysis and Management*, *5*, 340-362.

Greenberg, D., & Shroder, M. (2004). *The digest of social experiments, third edition*. Washington, DC: The Urban Institute Press.

Harvill, E. L., Peck, L. R., & Bell, S. H. (2013). On overfitting in analysis of symmetrically predicted endogenous subgroups from randomized experimental samples: Part three of a method note in three parts. *American Journal of Evaluation*, *34*, 545-556. doi: 10.1177/1098214013503201

Jacob, R. Zhu, P., & Bloom, H. S. (2009). New empirical evidence for the design of group randomized trials in education. New York, NY: MDRC. Available at <<http://www.mdrc.org/sites/default/files/new_empirical_evidence_fr.pdf >>

Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. *Theoria*, *69*, 301-321.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*, 604-20.

Laura and John Arnold Foundation. (2015). Low-Cost Randomized Controlled Trials to Drive Effective Social Spending. Available at << http://www.arnoldfoundation.org/wp-content/uploads/Request-for-Proposals-Low-Cost-RCT-FINAL.pdf>>

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, *25*, 17-38.

Meurer, W. J., & Lewis, R. J. (2015). Cluster randomized trials: evaluating treatments applied to groups. *JAMA Guide to Statistics and Methods, 313*, 2068-2069. doi:10.1001/jama.2015.5199.

Moulton, S., Peck, L. R. & Dillman, K. (2014). Moving to Opportunity's Impact on Health and Well-being Among High Dosage Participants. *Housing Policy Debate*, *24*, 415-446. doi: 10.1080/10511482.2013.875051

Nichols, A. (2012). *Evaluation of community-wide interventions*. Washington, DC: The Urban Institute. Available at

<<http://www.urban.org/research/publication/evaluation-community-wide-interventions>>

Olsen, R. B., Bell, S. H., & Nichols, A.. (2016). *Using Preferred Applicant Random Assignment (PARA) to Reduce Randomization Bias in Randomized Trials of Discretionary Programs*. Available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2850763

Olsen, R. B., & Orr, L. L. (2016). On the "where" of social experiments: Selecting more representative samples to inform policy. *New Directions for Evaluation*, *152*.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*, 107-121.

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

Orr, L. L., Bloom, H. S., Bell, S. H., Doolittle, F., Lin, W., & Cave, G. (1996). *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, DC: The Urban Institute Press.

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post treatment choice. *American Journal of Evaluation*, *24*, 157-187. doi: 10.1016/S1098-2140(03)00031-6

Peck, L. R. (2005). Using cluster analysis in program evaluation. *Evaluation Review, 29*, 178-196. doi: 10.1177/01933841X04266335

Peck, L. 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*, 256–274. doi: 10.1177/1098214007304129

Peck, L. R. (2013). On analysis of symmetrically predicted endogenous subgroups: Part one of a method note in three parts. *American Journal of Evaluation, 34*, 225–236. doi: 10.1177/1098214013481666

Peck, L. R. (2015). Using Impact Evaluation Tools to Unpack the Black Box and Learn What Works. *Journal of MultiDisciplinary Evaluation*, *11(24)*: 54-67.

Peck, L. R. (2016a). On the "how" of social experiments: Analytic strategies for getting inside the black box. In L.R. Peck (Ed.), Social experiments in practice: The what, why, when, where, and how of experimental design & analysis. *New Directions for Evaluation, 152,* 85–96.

Peck, L. R., ed. (2016b). Social Experiments in Practice: The What, Why, When, Where, and How of Experimental Design & Analysis. *New Directions for Evaluation, 152*. San Francisco, CA: Jossey-Bass/Wiley.

Peck, L. R., & Bell, S. H. (2014). *The role of program quality in determining Head Start's impact on child development* (OPRE Report #2014-10). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Peck, L. R., & Scott, Jr., R. J. (2005). Can welfare case management increase employment? Evidence from a pilot program evaluation. *Policy Studies Journal*, *33*, 509-533.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*, 463-479. doi: 10.3102/0162373709343964

Puma, M. J., Burstein, N. R., Merrell, K., & Silverstein, G. (1990). *Evaluation of the Food Stamp and employment and training program final report: Volume 1*. Bethesda, MD: Abt Associates.

Puma, M. J., Cook, R., Bell, S. H., Heid, C., Lopez, M., 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, M. J., Olsen, R. B., Price, C., & Bell, S. H. (2009). *What to do when data are missing in group randomized controlled trials*. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Available at <<http://ies.ed.gov/ncee/pdf/20090049.pdf>>

Reichardt, C. S. (2011). Evaluating methods for estimating program effects. *American Journal of Evaluation*, *32*, 246-272.

Schochet, P. Z., Burghardt, J., & Glazerman, S. (2001). *National Job Corps study: The impacts of job corps on participants' employment and related outcomes*. Princeton, NJ: Mathematica Policy Research.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York, NY: Houghton Mifflin.

Shadish, W. R. (2011). Randomized controlled studies and alternative designs in outcome

studies challenges and opportunities. *Research on Social Work Practice*, *21*, 636-643. doi: 10.1177/1049731511403324

Smith, J. (2002). *Evaluating Active Labor Market Policies: Lessons from North America*, unpublished manuscript. Available at <<http://www.glue.umd.edu/~jsmithz>>.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*. doi: 10.1177/1098214014527337

Tipton, E., Hedges, L. V., Vaden-Kiernan, M., Borman, G. D., Sullivan, K., & Caverly, S. (2014) Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*, 114-135.

Tipton, E. & Peck, L. R. (2016). A Design-based Approach to Improve External Validity in Welfare Policy Evaluation. *Evaluation Review*. doi: 10.1177/0193841X16655656