

This work is on a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license, <https://creativecommons.org/licenses/by-nc-nd/4.0/>. Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

1. INTRODUCTION

Many evaluations of social programs involve either multiple sites (a program implemented in more than one site at approximately the same time) or replication (one or more new sites adopting a program identical or similar to one that has previously been implemented), and sometimes both. Both approaches generate multiple tests of a program, which we refer to as “multiple trials.”

Although the two approaches are similar in important respects, there are also important differences, as multiple sites are implemented (roughly) simultaneously, while the implementation of replications is sequential. Multiple arm evaluations, in which programs that vary along certain elements but not others are tested at the same time on the same target population in the same sites, are a third type of multiple trial.

The purpose of this paper is to explore the rationales for evaluating programs and demonstrations through multiple trials, rather than limiting evaluations to a single trial, and to examine how such evaluations can best be designed to meet their objectives. After discussing the rationales for multiple trials, we develop a framework that allows us to categorize previous multiple trial evaluations according to the reason or reasons they were conducted and then assess whether their designs facilitated meeting their goals.

A key design concern we discuss is the extent to which treatments do and should vary across sites and between original programs and their replications. In addition, we discuss when multiple arms and subgroup analyses are superior to

multiple sites for accomplishing certain objectives. Key points are illustrated with examples from previous evaluations. Although these evaluations all relied on random assignment, many of the points also apply to evaluations that use non-experimental techniques.

While sets of multiple trials differ from one another in terms of whether the trials are simultaneous or sequential, they also differ along another dimension: how similar the individual trials are to one another. This may be viewed as a continuum. At one end are situations in which similar treatments are tested on similar target populations, but the environment differs because the sites, the time period, or both differ. Further along the continuum are multiple trials in which the treatments that are tested vary considerably, although they retain certain central elements. Moreover, the environments will likely differ, and the target populations may as well. Even further along the continuum, completely different treatments are tested, although the goals of the treatments tested are similar (e.g., helping low-wage workers retain employment), and the tests are conducted as part of the same project. We refer to the far end of the continuum as “multiple treatments.” Multiple treatments differ from multiple trials because the same program (or variants of it) is not tested. Of course, the point on the continuum that divides multiple trials from multiple treatments may be murky in practice.

All three types of multiple trial—multiple sites, replications, and multiple arms—are illustrated by the Greater Avenues for Independence (GAIN) evaluation of 1985. To assess GAIN prior to its widespread implementation, a

decision was made to use random assignment to conduct a multiple site evaluation in six of California's 57 counties. These six counties differed in numerous aspects, including the version of GAIN they implemented. For example, Riverside County emphasized moving members of the treatment group into the work force as quickly as possible, a so-called "work first" approach, while the other five counties relied to varying degrees more on a "human capital" approach in which education (especially adult basic education) and training (to a lesser extent) were used to upgrade the skills of welfare recipients *before* they sought work. It was hoped that by having multiple sites with differing program models, it would be possible to learn which was most effective. As it turned out, the program that operated in Riverside County was the most successful of the six (Riccio, Friedlander, & Freedman, 1994). This was widely attributed to Riverside's work first approach. However, because the counties differed in other important respects—for example, in terms of their economic characteristics and the characteristics of their AFDC recipients—definitive conclusions could not be drawn. In fact, the evaluators examined 14 explanatory variables, as well as interactions among these variables, in attempting to determine the reasons for the differences in impacts among the six counties resulting in an obvious degrees of freedom problem (Riccio & Friedlander, 1992, Chapter 6).

One way of drawing a more definitive conclusion is to run a replication to see if the work first approach can succeed elsewhere. Thus in 1995, Los Angeles County, which had been the least successful of the GAIN sites and which had

emphasized helping participants obtain remedial and basic education before seeking employment, adopted a modified version of the Riverside work first approach. An evaluation of this program, which used random assignment, found that it was far more successful than the original GAIN program in Los Angeles (Freedman, Knab, Gennetian, & Navarro, 2000). Although this finding is not definitive because the Los Angeles program did not entirely replicate the Riverside approach, and the two counties are obviously very different from one another, it was viewed as supportive of the work first approach.

2. THE RATIONALES FOR EVALUATING MULTIPLE TRIALS AND A FRAMEWORK FOR VIEWING THEM

Because it is usually costlier to conduct multiple trials than a single trial, one might ask what purposes are served by the former that cannot be served by the latter. In this section, we consider four. All but the first of these is in a sense a response to the fact that trials, even when they are part of the same evaluation, may vary in terms of the characteristics of their target populations (X), the economic and other environmental conditions at the sites and the time in which they are conducted (E), or components and inputs incorporated into the evaluated programs (P). Obviously, the resulting heterogeneity may cause program impacts to vary across the trials, as suggested by the following equation:

$$Y_{ijk} = \alpha + \sum \beta_i X_{ijk} + \sum \gamma_j P_{ijk} + \sum \delta_k E_{ijk} + \lambda Z_{ijk} + \sum \beta_{iz} Z_{ijk} X_{ijk} + \sum \gamma_{jz} Z_{ijk} P_{ijk} + \sum \delta_{kz} Z_{ijk} E_{ijk} + \epsilon_{ijk},$$

where Y is an outcome such as earnings; Z is the treatment variable (equal to 1 for those in the treatment group and 0 for those in the control group and thus does not vary in dosage); β , γ , δ , and λ are estimates of the respective effects of X , P , E , and Z ; i , j , and k respectively refer to a specific individual, program, and site; and ϵ is an error term.¹ The interaction terms allow for the possibility that program impacts may vary as X , P , and E vary. For example, a program may result in a larger increase in earnings at sites where the unemployment rate is low rather than high (or vice versa). If there are no interactions between Z and X , P , and E , then λ will capture the entire program impact.

a) Four Rationales for Multiple Trials²

(1) *Expanding the sample size.* A larger sample increases the possibility of statistically significant findings when the program has a non-zero impact and better allows for analyses of separate subgroups. In the case of replications, the sample drawn in an original trial sample may have been small because of cost considerations, but because of some hint of success with the original trial, a follow-up trial may be implemented with a larger sample. In the case of multiple trials, there may be relatively few persons in a tested program's target population at individual sites, but a sufficiently large sample can be obtained by testing the program in several sites. Possibly more important disruption can be minimized in implementing tests of modifications to ongoing programs by randomly assigning relatively few individuals in each of a number of sites to a control group, which is denied services, and then pooling the sample across the sites. Such an approach

was used for recent evaluations of the Job Corps and the Workforce Investment Act (WIA) and is currently being used in the ongoing evaluation of the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) program.

(2) *Finding the most effective program design.* As suggested by the GAIN evaluation, a multiple site experiment that tests programs that vary across the sites can help determine which program design is superior, but because the characteristics of the target population (X) and site environmental conditions (E) will likely also vary, this determination is inevitably subject to considerable uncertainty. Although this uncertainty can be reduced through replication, it cannot be eliminated. Thus, a better approach may be to run a multiple arm experiment.

In fact, this was ultimately done as part of a set of randomized experiments that took place after GAIN: The National Evaluation of Welfare-to-Work Strategies (NEWWS). To determine whether the work first approach really is superior to the human capital approach, welfare recipients at three NEWWS sites (Riverside; Atlanta, Georgia; and Grand Rapids, Michigan) were randomly assigned to either a control group, which received no services under the programs being evaluated, or to one of two different programs, each based on one of the two approaches. The two programs were then simultaneously operated at each of the three sites. The work first programs in all three sites tended to produce results that were superior to those relying on the human capital approach (Hamilton et al., 2001).³ Note, however, that a multiple arms

experiment, although clearly advantageous, became of interest only after it had been determined that there was a contest between program designs. This was not evident when the GAIN experiment was implemented. Thus, in the absence of enough foresight, multiple site evaluations may be all that is possible.

When there is great uncertainty as to the sort of program that can best ameliorate a particular social problem, rather than using a multiple trials approach, it may make more sense to use a multiple treatments approach in which P varies greatly across trials, although the goals of the different treatments are similar. The U.S. Department of Health and Human Services (DHHS) has made extensive use of this approach. Two examples are the Employment Retention and Advancement (ERA) project and Pathways for Advancing Careers and Education (PACE). In both instances, the Department started with an important issue they wanted addressed, but no single strategy appeared to be a likely solution. Thus, both ERA and PACE tested programs that differed vastly from one another. For ERA, the goal was to develop and test strategies to assist welfare recipients to keep their jobs and gain the skills needed to advance to higher paying jobs. The PACE projects were intended to develop strategies whereby low-income individuals could enter occupations where career pathways were available to enable them to gain additional skills and move on to better paying occupations. However, there were significant differences across sites. For example, the Des Moines Area Community College focused on individuals who lacked the basic skills required for the college's regular education and

training programs and offered short-term training and education with the goal of enabling participants to attend the regular vocational offerings. Year Up, another PACE component, focused on mature high school graduates who were not college bound and provided intensive vocational training and an internship to enable participants to increase their earnings at the end of the one-year experience (Fein & Hamadyk, 2018).

Although the multiple treatments approach does not lead to a single overall impact result, DHHS has treated such projects as integrated studies, and draws lessons from the entire group of studies considered together. For example, the ERA project includes 27 reports. Twelve of the reports are the individual site evaluations, but others are cross-cutting reports that compare the implementation and outcomes of alternative strategies used in the demonstration, as well as experiences of specific subgroups and specific strategies.⁴ Thus, although the multiple treatments approach is not a multiple trials approach as we have defined the term, it can be a powerful way to conduct demonstrations and impact evaluations when the situation calls for independent impact evaluations.

In the case of replication, sometimes the original program is tweaked in the replication to see if it can be improved (or operated at a lower cost). In this situation, the subsequent trials would ideally take place at the same site(s) as the original trial to hold X and E as constant as possible, but this has rarely been done and, in any event, such trials are still subject to changes in X and E over time. Same-site replication can be (and is) done to see if a seemingly successful

program can succeed when run at a larger scale. For example, sectorial training programs were originally tested in three sites and then replicated on a larger scale at four sites, one of which, Per Scholas in New York City, was common to both sets of trials (Hendra et al., 2016). However, as this example suggests, evaluations involving scale increases are typically done at a different site, even though this allows X and E to vary.

(3) *Increasing external validity.* With a sufficiently large sample, a one-time randomized experiment in a single site that has been properly implemented assures internal validity, but it provides no assurance of external validity. As implied by the equation shown at the beginning of this section, program impacts can vary because of heterogeneity in the characteristics of the target population (X) and in site environmental conditions (E). This heterogeneity is perhaps the major rationale for multiple trials. Even with heterogeneity, external validity can, in principle, be assured if the sites and subjects are both randomly drawn from the target population (see Olsen et al., 2013). This was done, for example, with the recently completed Benefit Offset National Demonstration (BOND) experiment, which had a sample of almost a million (968,530) disabled recipients of supplemental security income (SSI) or social security disability insurance (SSDI) and 10 sites that were randomly selected from the Social Security Administration's 53 Area Offices (Gubits et al., 2018). For analysis purposes, the sample was pooled across the sites.

In practice, however, randomly selecting sites has been relatively rare, in part because sites often have the option of whether to participate. Even if it is not done, some confidence can nonetheless be gained if a program is tested across sites that vary in terms of participant characteristics and environmental conditions, and thereby produces impact estimates that are more representative of an overall target population. For example, job clubs were tested in the 1970s and 1980s in a series of replication experiments, each in different sites with a different target population, that provided convincing evidence that they are an effective approach in assisting a variety of job seekers in finding employment. Due to these experiments, which are described in Section 4, the job club approach went from being considered an unorthodox way of helping job seekers to becoming the norm in a bit more than a decade.

(4) *Learning how various factors influence program impact.* In principle, multiple trials can be used to draw inferences about underlying “production” relationships by examining how program impacts vary with cross-site differences in X , E , and P . When X and P vary, this is sometimes characterized as “finding what works best for whom.” However, if the objective is to determine for whom a particular program works best, and P is relatively constant, this might be better accomplished by estimating impacts for separate subgroups (as categorized, for example, by gender, age, or attachment to the labor market).

When P does vary across trials, learning about production relationships would ideally involve estimating an equation like the one that appears at the

beginning of this section. This would be accomplished by using methods found in the literature on meta-analysis and hierarchical analysis. However, using these methods requires a large number of trials (Author, 2003), and few evaluations of social programs have a sufficient number (recall the degrees of freedom problem in the GAIN evaluation). One possible exception is the evaluation of the Food Stamp Employment and Training Program, which included 53 sites in 23 states. However, neither hierarchical nor meta-analysis was performed as these methods were rarely used by social scientists when the evaluation was conducted (around 1990). Nevertheless, there was an attempt to examine whether the site-level impacts differed by site characteristics by regressing these impacts against various site characteristics, but no statistically significant relationships were found (Puma et al., 1990). Very recently, the ongoing evaluation of the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) program evaluation, which has 88 sites, with roughly 60 families randomly assigned to the treatment group per site, used methods from hierarchical analysis to examine how impacts vary with program features and services received. However, while measures of P were examined, no variables that measured X or E were included (Michalopoulos et al., 2019).

In the absence of enough sites in one multiple site experiment, meta- or hierarchical analysis methods can still be used if data from several experiments can be pooled.⁵ The relation between program impacts and X, E, and P have, of course, also been less formally examined, as the GAIN example suggests.

One potential issue in determining the factors that cause impacts to vary in a multiple trial experiment is that the individual sites may have the ability to determine some of the components incorporated into the tested program and how the program is implemented. If so, it is possible that the factors that influence the choices in given sites are also related to the impact of the program and that some of these factors are difficult to measure and, hence, control for in the analysis. Thus, omitted variable or measurement error bias can result. Consider, for example, a multi-site evaluation of a vocational training program in which remedial education is also provided in some, but not all, the sites. If enthusiastic leaders are more likely to add a vocational training component and their fervor also results in larger program impacts, estimates of the contribution of vocational training to the overall program impact may be biased if, as is likely, leader enthusiasm is not measured and included in the analysis. Bell et al. (2016) have suggested a method that can be used to attempt to reduce this bias under certain circumstances.

b) A Multiple Trials Framework

Table 1 presents a framework based on the concepts discussed above. The table suggests that evaluations can be first be categorized as to whether they are of pilot or demonstration programs or of ongoing programs and then by the rationale for conducting a multiple, rather than single, trial evaluation. Of course, many evaluations have more than one rationale, which, as discussed later, may or may not be consistent with one another. For example, a goal of

expanding the sample size is consistent with the other three objectives listed in Table 1, but as indicated below, selecting the best program design may not be consistent with increasing external validity.

One reason for treating evaluations of pilot programs and evaluations of ongoing programs separately is that the latter almost always involve multiple site designs, drawing from sites where the program already exists, whereas the former may be based on replication or multiple arm designs, rather than multiple site designs. If the rationale for conducting an evaluation of a pilot program was either to expand the sample size or to increase external validity, either replication or multiple sites is an appropriate approach, but if it is to select the best program design, the ideal approach is to use multiple arms. Multiple arm experiments have several advantages for the latter purpose: the environment and participants characteristics are held constant, no more time to test treatments is required than in a multiple site evaluation, and additional sites are not required. As previously mentioned, however, multiple arm experiments may not be used in practice because of imperfect foresight, so replication or multiple sites may be relied on instead. Thus, the superiority of Riverside's work first approach was much less evident when tested in the multiple site GAIN then when tested as part of the NEWWS evaluation where a multiple arms approach was used.

TABLE 1
Preferred Multiple Trials Strategies for Alternative Learning Objectives

Program Type	Pilot				Ongoing			
	Expanding sample size	Selecting the best program design	Increasing external validity	Studying heterogeneity in X, E, & P	Expanding sample size	Selecting program design	Increasing external validity	Studying heterogeneity in X, E, & P
<i>Preferred evaluation design</i>								
Type of multiple trial	MS or R	MA	MS or R	MS	MS	MS	MS	MS
Heterogeneity Permitted in								
Program (P)	NR	Allowed	Not allowed	Allowed	NR	Allowed	Allowed	Allowed
Participants (X) & Environment (E)	NR	Not allowed	Allowed	Allowed	NR	Allowed	Allowed	Allowed

Notes: MS= multiple sites; R= replication; MA=multiple arms, NR=not relevant

A second reason for distinguishing between evaluations of pilot programs and demonstrations and evaluations of ongoing programs is suggested by a literature in public health that deals with program heterogeneity in terms of efficacy and effectiveness trials. In efficacy trials, a treatment is evaluated in terms of optimum participation and implementation, and often at a small scale. Ideally, but not always in practice, these are then followed by treatment effectiveness trials, which evaluate the program in a “real world” setting, often increasing the scale of operations. These include implementation trials in which program variation is permitted and program evaluation trials in which variation in both participant behavior and implementation are allowed (see Gottfredson et al., 2015). Although these terms are not usually used in discussing evaluations of employment, training, and education programs, in the case of pilot programs, it is

often efficacy trials, rather than effectiveness trials, that are actually evaluated. This can occur, for example, if programs are tested in the sites that are most likely to administer a treatment successfully, the individuals selected into treatment are those most likely to benefit from the treatment, the program was optimized for the conditions existing in the selected sites, or intensive technical assistance that would not exist in an ongoing program is provided to the sites (see Banerjee et al. (2017) for a discussion of these factors). While evaluations of many, if not most, pilot programs and demonstrations are efficacy trials, almost by definition, evaluations of ongoing programs are typically effectiveness trials.

The bottom two rows of Table 1 indicate whether the evaluated program design and participant and environmental characteristics should ideally be held constant across trials or allowed to vary. As mentioned in the previous paragraph, multiple site evaluations of ongoing programs are usually efficiency trials in which cross-site variation in program designs and participant and environmental conditions is almost inevitable. Such variation is not necessarily true of evaluations of pilot programs. For example, the intention of a multiple arm experiment is to assess alternative program designs by holding participant and environmental characteristics constant while allowing program design to vary across trials. In contrast, if the goal of a pilot program evaluation is to increase external validity, variation in participant and environmental characteristics is critical, while the program design would ideally be kept constant across the trials, although, as discussed below, holding the program design constant may not be

possible in practice. Of course, if the goal of a multiple trial evaluation is to study heterogeneity in program design and participant and environmental conditions, all these factors must be allowed to vary regardless of whether the evaluation is of ongoing or pilot programs.

As is apparent from Table 1, it is very difficult to design a single experiment that both maximizes external validity and permits selection of the best program design. As designed, for example, the GAIN experiment did not fully accomplish either goal, although the intent was to address both. By itself, the six-site GAIN evaluation pointed to the work first approach as being the best of the tested treatments, but not definitively, and it provided little information about the external validity of the Riverside results.

If the intent of an evaluation is to either study heterogeneity in program design or select the best program design, it is reasonable to estimate a multiple treatment model in which separate impacts are estimated for each trial, and the program is assessed on a trial-by-trial basis. Thus, the multiple treatments model was emphasized in the GAIN evaluation because there were important differences among the programs run in the sites, and there was an interest in determining which program model was most effective. In contrast, if the goal of an evaluation is to determine whether a program is effective overall, a common impact model in which the impacts are averaged across the sites is often estimated. The common impact model is useful in a multiple site study when the sites are considered part of a single program, which is typically the case with

ongoing programs and sometimes true of pilot programs, and together the sites are considered more or less representative of the program's target population and service mix. For example, the common impact model was emphasized in the evaluation of the Food Stamp Employment and Training Program, an ongoing program, even though there was considerable variation in what the different sites offered. In this case, which is discussed further in the following section, the sites were randomly selected, and the goal of the evaluation was to determine whether the overall program, as it actually existed in the field, was effective.

3. USING THE FRAMEWORK TO CATEGORIZE AND ASSESS PREVIOUS EVALUATIONS

As implied by Table 1, previous multiple trial evaluations can be categorized by whether they are of pilot programs or ongoing programs and by the rationale for using a multiple trial approach. The specific design that was adopted can then be assessed. This is illustrated in this section for the three arguably most important categories that appear in Table 1: (a) evaluations of pilot programs in which multiple trials are intended to aid in selecting a program design; (b) evaluations of pilot programs in which multiple trials are intended to increase external validity; and (c) evaluations of ongoing programs in which multiple trials are intended to increase external validity.

(a) Evaluations of pilot programs intended to aid in selecting a program design

As discussed in Section 2, the impacts of variants of a program containing core elements that are tested in different sites can be compared to select the

variant that appears most effective. The multiple site GAIN evaluation provides one example. A second example is a series of randomized experiments targeted at Unemployment Insurance (UI) claimants. The first of these experiments, which was conducted in 1983 in Charleston, South Carolina, required UI claimants to participate in job search activities and undergo stronger enforcement of the UI work test that they be available and actively searching for work (Corson et al., 1984). Subsequent experiments in Washington State and New Jersey tested alternative combinations of reemployment services and work test enforcement that differed from those tested in Charleston (Johnson & Klepinger, 1991 and Corson & Haimson, 1996).

Because sites differ in participant and environmental characteristics, as pointed out in Section 2, with sufficient foresight, testing program variants can often be better done with multiple arms. As Peck (2020) notes, multiple arm experiments “can compare distinct program models to one another (the ‘competing treatments’ design) or they can compare alternative versions of the same program model (the ‘enhanced treatment’ design).” The direct comparison of the work first and human capital approaches in the NEWWS evaluation is one example of the former, but there are many others. The Minnesota Family Investment Program (MFIP) demonstration, which took place in the mid-1990s, is an example of the latter. In this experiment, AFDC recipients and applicants assigned to both treatment arms were eligible for financial incentives

encouraging employment, but individuals in only one arm were subject to mandatory participation in employment services (Miller et al., 2000).

Multiple arm experiments can accommodate a wide variety of specific designs.⁶ The key element, as noted by Orr (2020), is that the experiment “randomly assigns sample members from a common pool to all treatment groups and the control group.” The ultimate multiple arm experiments are those designed to estimate “response surfaces”—that is, their designs allowed program parameters to vary within wide ranges. For example, eight combinations of income guarantees and tax rates were tested in the New Jersey Income Maintenance Experiment, and over 40 combinations were tested in the Seattle-Denver Income Maintenance Experiment. The RAND Health Insurance Study and the Housing Allowance Demand Experiment also used response surface designs.⁷

(b) Evaluations of pilot programs intended to increase external validity

Either multiple sites or replication can be used to determine whether a program is effective in different settings.⁸ The important thing is that the program tested in each trial be as similar as possible or at least not differ beyond relatively modest adjustments to allow for differences in local conditions or in target populations, but participant and environmental characteristics vary. If there is interest in varying business cycle conditions or to see if a program continues to work if tested at a larger scale, it may be necessary to conduct replications at different times. Such replications could, in principle, be conducted within the

same site. In fact, it may be best to do so in order to hold other factors constant. As previously mentioned, however, in practice, most replications involve the use of new locations.

Banerjee et al. (2017) provide a useful discussion of using replication to test pilot programs on a larger scale. They first discuss “six challenges” in moving from small-scale efficacy trials to “a policy implemented at scale: market equilibrium effects, spillovers, political reactions, context dependence, randomization or site-selection bias, and implementation challenges.” They then illustrate some of the challenges with an educational intervention in India that was first tested through a small-scale efficacy trial and then through a series of large-scale efficiency replications that involved finding the best way to implement the intervention. The intervention has now been widely adopted in India.

The work of Azrin and his colleagues in testing the efficacy of job clubs is an important example of using replications as a program validity test under varying conditions. Azrin challenged the conventional wisdom that job seekers should be served individually by first introducing group activities in what he called “job clubs” for a small number of job seekers in a small community; then he replicated the work with other groups.

The first job club experiment involved a sample of 120 job seekers from a small college town in southern Illinois (Azrin, Flores, and Kaplan, 1975). The job club treatment included the then-novel feature of group counseling, but the program also emphasized that job search should be a full-time job, and included

a “buddy system,” motivational talks and information, family involvement, encouragement to consider a wide variety of jobs, having participants teach the curriculum to new members of the group, and instruction on dress and grooming. Based on a final analysis sample of 60 individuals who were randomly assigned to the treatment and control groups, the evaluation findings were impressive. The median treatment group member started work in 14 days, while the median control group member started work in 53 days. Three months after random assignment, 92 percent of the treatment group had found a job, compared to 60 percent for the control group. Average wage rates for those who found a full-time job were 36 percent higher for the treatment group. All the impact findings were statistically significant with $p < .05$.

Azrin and colleagues replicated their work with two specific populations. Azrin and Philip (1979) used a randomized field experiment to test the job club approach for a group of 154 job seekers with disabilities. The results were again impressive. Six months after random assignment, 95 percent of the treatment group had obtained a job, compared with 28 percent of the control group. With support from the U.S. Department of Labor, Azrin and colleagues were then able to conduct a much larger randomized field experiment with recipients of AFDC in five cities. A total of 967 clients were randomly assigned to either treatment or control status, with the treatment group receiving services similar to the model in Azrin’s other work and the control group receiving the standard services provided to welfare recipients in those sites. For the entire sample at the 12-month follow-

up, 87 percent of the treatment group had obtained jobs compared to 59 percent of the control group. There were statistically significant differences favoring the treatment group for each site and virtually all the subgroups considered.

The final replication of the job club model was performed by the Manpower Demonstration Research Corporation (now MDRC) in a Louisville, Kentucky welfare office (Wolfhagen & Goldman, 1983). Once again, the evaluation found strong employment results for the treatment group. Of the 750 welfare recipients who registered in Louisville between October 1980 and May 1981 and who participated in the demonstration, average earnings for the treatment group over this period were \$550, as compared to \$144 for the control group.

The Azrin job club studies and the MDRC replication provide convincing evidence that job clubs are an effective approach in assisting a variety of job seekers in finding employment. The job club approach went from being considered an unorthodox way of helping job seekers to becoming the norm in just over a decade, in large part due to the efforts by Azrin and his colleagues. As suggested by the widespread adoption of job clubs, replications of experiments can generate considerable enthusiasm for a policy, at least when the findings are consistent across the trials for a variety of target groups.

Rather than taking Azrin's approach of sequentially testing the same program on different subgroups, however, if the groups of interest are available in a study sample, then a superior approach is to use standard subgroup

analyses to assess if a program is successful for different groups. In addition, subgroup analysis can be (and is) used to determine whether treatment effects vary across groups.

As indicated above, if the intent of an evaluation of pilot programs and demonstrations is to increase external validity, then it is important to minimize variation in the program model across trials, as was done in the job club replications. This was also accomplished in the multiple site Maryland Primary Prevention Initiative (PPI), which operated from 1992-97 in six welfare offices in Maryland and was similar across the sites (see Stoker & Wilson, 1998).

Sometimes, however, there is a need to adapt to differences in local conditions or target populations. For example, job openings for different occupations may vary across sites. Year Up provides an interesting example.⁹ The tested program focuses on high school graduates, aged 18-24, and operates in nine locations across the country. All Year Up programs last one year, with the first half focusing on developing occupational and soft skills, and the second half spent in an internship at a major corporation or nonprofit organization. Although the basic structure of the program was consistent across sites, the national office did not provide standard curricula, and course content varied among sites. The ongoing evaluation (Fein & Hamadyk, 2018) found that the Year Up sites maintained strong allegiance to the program model, largely through the efforts of the national office.

If a multiple site evaluation is used to determine the external validity of estimates of the impacts of a treatment, it is important that the sites be as representative as possible of the overall target population and the local environments in which they reside. For example, the GAIN evaluation included six sites that represented a diverse set of local economic conditions and population characteristics within California. Los Angeles County had to be included because one-third of the AFDC caseload lived there. In addition, there were three other urban counties (Riverside, San Diego, and Alameda, which includes Oakland), and two rural agricultural counties (Tulare and Butte, both in the Central Valley, with Butte in the northern part and Tulare in the southern part of the 450-mile long valley). Unfortunately, although the sites represented a wide range of conditions within California, the evaluation was of limited value in determining the external validity of a specific program because what was defined as “GAIN” differed in important respects across the sites.

While it may be necessary to adopt a program design to local conditions or differences in target populations, sometimes excess variation in a tested program occurs because of a lack of fidelity to the program design, although it is often difficult, in practice, to distinguish purposeful variation from variation due to a lack of fidelity.¹⁰ A clear case of the latter occurred in the Center for Employment Training (CET) replication. A high degree of fidelity appears to have only been approximated in four of the 12 CET-replication sites. Among the

remaining sites, six implemented the program with medium fidelity and the other two with low fidelity (Miller et al., 2003).

As this example suggests, good fidelity is frequently difficult to attain in field experiments in practice because it is often challenging to get sites to follow an exact formula. One possible approach to minimize poor fidelity is to obtain a pre-experimental agreement from each site to implement the program model in accordance with detailed specifications before including them in the experiment. Once the experiment is underway, process analysis is important for monitoring fidelity. Consideration should also be given to developing a check list for social experiments along the lines of the “replication recipe” that Brandt et al. (2014) constructed for experimental social psychology. Such a recipe could be used for multiple site experiments, as well as for replications.

(c) Evaluations of ongoing programs intended to increase external validity

Some ongoing programs permit or even encourage considerable local or regional variation. If the goal is to evaluate a program as it exists, then each site should operate the program as they typically do. In other words, treatment effectiveness evaluations are appropriate. As discussed earlier, this is best accomplished through a representative sample of sites, ideally by randomly selecting sites out of the potential population of sites (see Olsen et al., 2013). Examples of this are the decades old evaluation of the Food Stamp Employment and Training Program (Puma et al., 1990), which appears to be the first randomized experiment to draw a nationally representative sample of sites that

mimics how a national program is administered locally, and the recent evaluations of the Workforce Investment Act (Heinrich et al., 2013; Fortson et al., 2018). These evaluations covered programs that varied considerably across states and across local areas.

In practice, however, random assignment of sites has rarely been done, in part because, under many programs, sites can decide for themselves whether to participate in an evaluation. The evaluation of the Food Stamp Employment and Training Program was an important early exception because random assignment could be imposed by the federal government.

For an ongoing program in which sites have a great deal of discretion in how they operate the program, policy makers likely are most interested in the overall impact of the program as it actually operates, although learning about the impacts of specific components, dosages, and strategies can also be useful for program improvement.¹¹ In so-called “block grant” programs, such as the Workforce Innovation and Opportunity Act (WIOA), grant recipients (often state and local governments but sometimes winners in competitions) often must follow some rules on the services they provide, but they are also likely to have a great deal of flexibility in how they serve participants. We illustrate the evaluation issues by discussing WIOA and its predecessors below.

The national evaluations of WIOA’s predecessors, the Job Training Partnership Act (JTPA) and the Workforce Investment Act (WIA) both dealt with programs that were highly diverse among states and local areas, where the

diversity was an important aspect of the program. Under all three programs, federal funds were allocated to states by formulas based on indicators of population and economic distress, and similar formulas were used to distribute the money to local areas within states. States were permitted to strongly influence the mix of workforce activities in the state, and many states passed on this discretion to local areas. To illustrate: under WIA, activities were categorized as core, intensive, or training, and participants were to receive the services sequentially. Core activities were the least expensive, and typically involved providing labor market and job search information to participants, often through self-service or in a group setting. Intensive services involved more one-one-one assistance, often as in-depth counseling or assessment. Finally, training involved vocational training typically provided by community colleges, nonprofit institutions, or proprietary schools.

Training programs typically cost thousands of dollars, while the assistance provided in core and intensive services costs hundreds of dollars at the most. Because training programs are intended to increase human capital and thereby earnings, training programs would be expected to increase earnings, at least in the long run, by significantly more than programs focusing on job search assistance. Author (2007) found that there was a very large range in the proportion of participants who received training among states. For the 2002-2005 period nationwide, 46.6 percent of WIA adults who exited the program received training. The range, however, was from 96 percent in Delaware, down to 13.5

percent in Mississippi. With such strongly divergent approaches to implementing their programs, program impacts on earnings could vary significantly among states and local areas.

The JTPA and WIA evaluations dealt differently with the expectation that training was likely to have a different impact than other services. In the JTPA evaluation, sites determined the most appropriate service for each person they wished to serve, and then individuals were randomly assigned to treatment or control status. In the WIA evaluation, individuals were randomly allocated to be eligible for only core services, only core plus intensive services, or all three tiers of service; no one was assigned to a pure control service where they were ineligible for any services.

To gauge the effectiveness of the overall program, the national evaluations of JTPA and WIA used the common impact model approach, which computes the average impact across all sites.¹² The National JTPA Study included 16 sites, but only those that were willing to participate. Although the authors argued that the sample was representative of the nation, others expressed concern that the sites were not selected randomly so that the impact in the participating sites might differ from that in a random sample.¹³ The design for the WIA evaluation overcame these problems. The sample included 28 randomly selected sites, and only two sites were replacements for sites that refused to participate. Neither the JTPA evaluation nor the WIA evaluation

reported site-specific impacts, but both evaluations investigated if activities were related to impacts.

4. CONCLUSIONS

Conducting evaluations with multiple trials involves added costs, but doing so can improve our knowledge of program impacts in several ways. In Section 2 of the paper, we note that adding additional sites provides several advantages for the evaluation. First, adding additional trials means that sample size will be increased, thus increasing the probability of obtaining statistically significant findings when there is a non-zero impact. Second, adding additional sites can be useful in identifying the most effective program design. Adding sites enables the evaluation to identify variations in the target population, the program design, and the environment that affect the outcomes of interest. However, in some cases, multiple arms provide a superior way of determining the best program design. Third, adding more sites can increase external validity by incorporating greater and more representative variation in population characteristics, program characteristics, and the environment. Finally, conducting multiple trials enables the evaluation to explore how variations in population, program, and environment affect the outcomes. Subgroup analysis, however, often also plays an important role in examining how program impacts are affected by population variation. As indicated by the framework introduced in Table 1 and illustrated in Section 3, the extent to which each of these goals is emphasized and whether a pilot program

or ongoing program is involved should influence the specific multiple trial evaluation design that is selected.

REFERENCES

- Azrin, N. H., Flores, T., & Kaplan, S.J. (1975). Job-Finding Club: A group-assisted program for finding employment. *Behavioral Research and Therapy* 13, 17-27.
- Azrin, N. H. & Philip, R.A. (1979). The job club method for the handicapped: A comparative outcome study. *Rehabilitation Counseling Bulletin* 23, 144-155.
- Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., ... Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives* 31, 73-102.
- Banerjee, A., Banerji, R., Berry, J., Duflo, E., Glennerster, R., & Khemani, S. (2008) Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. Cambridge, MA: Massachusetts Institute of Technology.
- Author, 2010
- Author, 2007
- Bell, S, Harvill, E. L., Moulton, S. R., & Peck, L. R. (2016). *Can evaluators reduce cross-site attributional bias in connecting program elements to program*

impacts using within-site experimental evidence? Bethesda MD: Abt Associates.

Bell, S. H. & Peck, L. R. (2016). On the “how” of social experiments:

Experimental designs for getting inside the black box. L.R. Peck (Ed.) *The what, why, when, and how of experimental design & analysis. New Directions for Evaluations* 152, 97-107.

Bloom, H. S., Hill, C. J., & Riccio, J. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management* 22, 551-575.

Brandt, M. J., Jizerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R. ... van'tVeer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology* 50, 217-224.

Cook, T.D. (2014). Generalizing causal knowledge in the policy sciences:

External validity as a task of both multiattribute representation and multiattribute extrapolation. *Journal of Policy Analysis and Management* 33, 527-536.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Designs and Analysis Issues for Field Settings*. Chicago: Rand-McNally.

Corson, W. & Haimson, J. (1996). *The New Jersey Unemployment Insurance Reemployment Demonstration Project*. Washington, DC: U.S. Department

of Labor, Employment and Training Administration, Unemployment Insurance occasional paper 95-2, revised edition.

Corson, W., Long, D.A., & Nicolson (1984). *Evaluation of the Charleston Claimant Placement and Work Test*. Princeton NJ: Mathematica Policy Research.

Cronbach, L.J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Fein, D. & Hamadyk, J. (2018). *Bridging the opportunity divide for low-income youth: Implementation and early impacts of the Year Up program*. Bethesda, Maryland: Abt Associates, Inc.

Fortson, K., Rotz, D., Burklander, P., Mastri, A., Schochet, P., Rosenberg, L., & D'Amico, R. (2018). Providing public workforce services to job seekers: 30-month impact findings on the WIA Adult and Dislocated Worker Programs. Washington, DC: U.S. Department of Labor, Employment and Training Administration, occasional paper 2018-4_1.

Freedman, S., Knab, T., Gennetian, L. A., & Navarro, D. (2000). *The Los Angeles Jobs-First GAIN Evaluation: Final report on a work first program in a major urban center*, New York: Manpower Demonstration and Research Corporation.

Gottfredson, D.C., Cook, T.D., Gardner, F.E., Gorman-Smith, D., Howe, G.W., Sandler, I.N., & Zafft, K.M. (2015). Standards of evidence for efficacy,

effectiveness, and scale-up research in prevention science: Next generation.
Prevention Science. 16, 893-926.

Author, 1990

Author, 2003

Author, 2004

Author, 2005

Author, 2003

Gubits, D., Stapleton, D., Bell, S., Wood, M., Hoffman, D., Croake, S., Mann, D.

R Judkins, D. (2018) *BOND implementation and evaluation: Final evaluation report*. Washington, DC: Abt Associates.

Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J.,

Adams-Ciardullo, D., & Gassman-Pines, A. (2001). *National Evaluation of Welfare-to-Work Strategies: How effective are different welfare-to-work approaches? Five-year adult and child impacts for eleven programs*. New York: Manpower Demonstration Research Corporation.

Heinrich, C.J., Mueser, P.R., Troske, K.R., Jeon, K., & Kahvecioglu, D.C.

(2013). Do public employment and training programs work? *IZA Journal of Labor Economics*. 2, 1-23.

Hendra, R., Greenberg, D. H., Hamilton, G. A., Oppenheim, A. Pennington, A.,

Schaberg, K., & Tessler, B. L. (2016). *Encouraging evidence on a sector-*

focused advancement strategy: Two-year impacts from the WorkAdvance Demonstration. New York: MDRC.

Johnson, T. R. & Klepinger, D. H. (1991). *Evaluation of the impacts of the Washington Alternative Work Search Experiment*. Washington, DC: U.S. Department of Labor, Employment and Training Administration, Unemployment Insurance occasional paper 91-4.

Michalopoulos, C., Faucetta, K., Hill, C. J., Portilla, X. A., Burrell, L., Lee, H., Duggan, A., & Knox, V. (2019). *Impacts on family outcomes of evidence-based early childhood home visiting: Results from the Mother and Infant Home Visiting Program Evaluation*. New York: MDRC.

Miller, C., Bos, J. M., Porter, K. E., Tseng, F. M., Doolittle, F. C., Tanguay, D. N., & Vencil, M. P. (2003). *Thirty-month findings from the Center for Employment Training replication sites*. New York: MDRC.

Miller, C., Knox, V., Gennetian, L.A., Dodoo, M., Hunter, J., & Redcross, C. (2000). *Final report on the Minnesota Family Investment Program*. New York: MDRC.

Olsen, R.B., Orr, 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. (2020). Multi-Arm Trials for Informing Policy and Practice: A 50-Year Retrospective, unpublished paper.

Orr, L. L., Bloom, H. S., Bell, H. H., Doolittle, F., & Lin, W. (1996). *Does training for the disadvantaged work? Evidence from the National JTPA Study*.

Washington, DC, Urban Institute Press.

Peck, L.R. (2020). *Experimental Evaluation Design for Program Improvement*.

Los Angeles: Sage.

Puma, M. J., Burstein, N. R., Merrell, K., & Silverstein, G. (1990). *Evaluation of the Food Stamp Employment and Training Program: Final report*, Bethesda,

MD: Abt Associates.

Riccio, J., Friedlander, D., & Freedman, S. (1994). *GAIN: Benefits, costs, and three-year impacts of a welfare-to-work program*. New York: MDRC.

Riccio, J. & Friedlander, D. (1992). *GAIN program strategies, participation patterns, and first-year impacts in six counties*. New York: MDRC.

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

Stoker, R. & Wilson, L. (1998). Verifying compliance: Social regulation and welfare reform. *Public Administration Review*. 58, 395-416.

Author, 2007

Weiss, M.J., Bloom, H.S., & Brock, T. (2013). A conceptual framework for studying the sources of variation in program effects. New York: MDRC.

Wolfhagen, C. F. & Goldman, B. S. (1983). *Job search strategies: Lessons from the Louisville WIN Laboratory*. New York: Manpower Demonstration Research Corporation.

ENDNOTES

¹ Others have also used similar frameworks for discussing the factors that influence program effects. For example, Author (1990) present the following equation, which represents the production function for welfare-to-work programs: $Q = f(P, C, L)$, where Q is the outcome (our Y), P is program characteristics, C is client characteristics (our X), and L is labor market conditions (similar to our E). Our equation is also very similar to the framework developed in Cronbach (1982) and extended in Cook (2014). Cronbach (1982) describes the key aspects of an evaluation as “UTOS,” where U are units of analysis (e.g., persons, schools, etc.), T is the treatment to be provided (or withheld from the control group), O represents the outcomes or variables hypothesized to be affected by the treatment, and S represents the settings in which the intervention takes place. When written in lower case, utos refers to the values of these characteristics for the sample selected, while the uppercase UTOS refers to the universe. When the evaluator wishes to apply the findings to another set of values of U, T, O, and S, a * precedes the domain characteristics, i.e., *U, *T, *O, and *S. Cook (2014) interprets Cornbach’s framework and adds time (ti) to the model. In our equation, the X variables are similar to U, the treatment is Z rather than T, the outcome (O)

is Y, and the setting variables are the E or environment variables. Our model is simpler in that we include a single O variable and assume a linear functional form. We do not include time explicitly because we assume that differences in the outcome over time are due to changes in the X, T, or E variables rather than changes in time itself. These issues are also addressed in Weiss, Bloom, and Brock (2013).

² The four rationales presented in this section map well with three of the four validity concepts originally developed by Cook & Campbell (1979) and updated in Shadish, Cook, & Campbell (2002). Expanding sample size is an important way to increase statistical power and thereby increase statistical conclusion validity. The second and fourth rationales presented in this section, determining the best program design and determining how characteristics of the participants and environment affect impact, are related to internal validity. Shadish, Cook, & Campbell (2002, p. 38) define internal validity as “The validity of inferences about whether observed correlation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.” Evaluations focusing on whether characteristics of participants or the environment affect the program impact essentially modify the concept of internal validity, looking at the relationship conditional on the value of participant or environment characteristics. The third rationale in the paper, increasing external validity, explicitly considers one of the four types of validities discussed by Cook & Campbell (1979) and Shadish, Cook, & Campbell (2002).

The only type of validity they discuss that is not related to the section is construct validity, which deals with measurement of characteristics.

³ In interpreting these findings, it is important to recognize that the human capital arms in the three sites did not provide intensive investments in human capital. The human capital arms in Riverside and Atlanta mainly provided basic education; while the arm in Grand Rapids emphasized basic education, it also encouraged vocational training. The most successful of all the programs tested in NEWWS by far was one operated in Portland, Oregon that mixed the work first and human capital approaches. Like the former, this program emphasized to participants that the goal for them was to obtain a job; but rather than being told to take any available job immediately, participants were told to wait until they could find a “good” job. In addition, as in the human capital approach, those participants who needed additional skills were encouraged to enroll in education and training (Author, 2005).

⁴ <https://www.acf.hhs.gov/opre/research/project/employment-retention-and-advancement-project-era-1998-2011> retrieved March 16, 2019.

⁵Examples of when this has been done include Bloom, Hill, and Riccio, 2003; Author, 2005; and Author, 2003.

⁶ These are described in some detail by Peck (2020), Bell & Peck (2016), and Orr (2020), who also provide numerous examples of each type of multiple arm experiment. The units that are randomly assigned to the arms are not necessarily limited to individuals. For example, an experiment described in Banerjee et al.

(2008) randomly assigned 280 villages in India to four arms including the control group.

⁷ These experiments were very expensive, and they all took place in the 1960s and 1970s. A response surface design is only possible when the program parameters are continuous (e.g., tax rates or copayment rates). For brief descriptions of experiments using response surface designs, see Author (2004). Cook (2014, p. 532) also addresses response surface designs and refers to the approach as “probably the best available method for causal extrapolation,” although he is not optimistic about the potential for widespread use of the technique in the near term for the social sciences.

⁸ Because either multiple sites or replications may be used to examine external validity, it is useful to consider the trade-offs between them. Replications can hold the socioeconomic environment and participant characteristics constant by being run in the same site as the original experiment, although this seems to be done relatively rarely. More importantly, because replications are sequential, they require more time to test the same treatment. On the other hand, randomized multiple site experiments require an upfront investment to include additional sites. Decisions about whether to replicate an experiment, in contrast, can be based on what was learned from the initial evaluation. For example, if the initial treatment did not have positive impacts, there may be little point in fielding a replication. This can result in considerable cost savings. If positive impacts did occur, it may be deemed important to determine if they will hold up in a different

setting. Moreover, the initial evaluation may suggest variations in the treatment that appear promising because they potentially can result in even larger impacts or in lower costs. These can be tested in the second round. A further advantage of replication relative to multiple sites is that replications provide an opportunity to improve the evaluation methodology, either by using stronger methods such as a randomized controlled trial (RCT), rather than a comparison group, or by using a larger sample. For example, the I-BEST replication, which is currently being conducted, substituted a randomized control trial for non-experimental methods. In the Year Up evaluation (Fein & Hamadyk 2018), a randomized control trial was used in both the original and the replication evaluations, but the replication included eight sites rather than just three. Thus, the information provided by testing treatments through a sequential replication process can have considerable value. This value, however, must be compared to the value of the time required for the replication process to be completed. For example, as discussed below, around a decade elapsed between the first job club experiment in Carbondale, Illinois and the evaluation report on the final experiments in Louisville, Kentucky. However, the job club experiments demonstrate the feasibility of testing an innovation on a very small and inexpensive scale, finding success, and then ultimately testing it on a larger scale.

⁹ Fein & Hamadyk (2018) describe the Year Up program and the randomized evaluation of it that is in progress.

¹⁰ Another possibility is that the original design is silent on specific features, which would likely lead to variation among sites.

¹¹ A problem can occur if an attempt is made to extract information about what works best from the observed variation in ongoing programs. If the local adaptations are optimal, then if site A were to implement site B's program, it would have worse outcomes than under its previous program.

¹² See Orr et al. (1996) for the national JTPA evaluation and Fortson et al. (2018) for the national WIA evaluation.

¹³ Author (2010) discusses the problem with the sites in the National JTPA Study.