Sample Design

SOCIAL EXPERIMENTATION:
EVALUATING PUBLIC PROGRAMS WITH EXPERIMENTAL METHODS **PART 4**

Estimates of program effectiveness are just that — estimates. They are facts, not truth, and once we have them we are still left with the problem of deciding what, if anything, they tell us about truth and whether what they tell us is enough to guide our actions. Conversely, in setting out to collect facts, we want to design evaluations so that their results will be a useful guide to action once we get them.

— Stephen D. Kennedy, Chief Social Scientist, Abt Associates

This is the fourth in a series of papers on social experimentation. The first three papers focused on the rationale for social experiments and the design and interpretation of the experimental comparison. In this paper, we discuss the issues involved in designing the experimental sample to achieve the most valid and precise estimates of the experimental impact. Specifically, we address the following issues:

- Site selection and the external validity of the experimental impact estimates;
- Sample size and the statistical power of the design;
- The point in the sample intake process at which random assignment is conducted and the power of the design;
- Allocation of the sample among multiple treatments;
- The optimal number of experimental sites; and,
- Random assignment of groups of individuals.

Site Selection and the External Validity of the Estimates

As noted in the first paper in this series, experimental estimates are **internally valid**—that is, they provide unbiased estimates of the impact of the experimental treatment on the population to which it was applied. For the experimental estimates to be **externally valid**, the experimental sample must accurately represent the population of interest for policy. External validity is sometimes characterized as **generalizability**. In order to provide the most reliable guidance for policymakers, experiments should be both internally and externally valid. In designing experiments, then, it is important to pay careful attention to a number of threats to the external validity of the estimates.

Ideally, the experimental sample would be a random sample of the population of interest for policy.[1] Just as random assignment creates two groups that do not differ systematically in any way, random selection of the experimental sample from the broader population of interest would produce a sample that does not differ systematically from that population. Thus, if the experimental sample is a random sample of the population of interest, the impact estimates are unbiased estimates of what the impact of the program would be in the larger population.

In most applications, however, simple random sampling from the population of interest is not feasible. It would probably not be possible, for example, to conduct an experiment with a simple random sample of all AFDC recipients in the U.S., or even in a single state. Such a sample would be spread so thinly over a large number of program offices and geographic areas that the costs of experimental administration and data collection would be prohibitive. Instead, experimental samples are generally clustered in a small number of sites.

It is still possible to obtain a random sample of the overall population (*e.g.*, all AFDC recipients in the U.S. or in a given state) if the experimental sites are randomly selected from all sites in that population and experimental participants are randomly selected from the population of interest within each site. Such a sample design is known as a **multistage random sample**. This type of sampling procedure was used in the national evaluation of the Food Stamp Employment and Training Program (FSETP).[2] In that experimental study, a sample of 60 local food stamp agencies (FSAs) was randomly selected from among 410 FSAs containing 85 percent of the national population of program participants.[3] An intensive site recruiting effort

[1] Random sampling should not be confused with random assignment. In random sampling, a group of individuals is randomly selected from a larger population in order to obtain a sample for analysis that is representative of the population from which it was drawn. In random assignment, the analysis sample is randomly divided into two or more groups to be subjected to different policy regimes.

[2] See Puma et al. (1990).

[3] FSAs serving less than 50 participants per year were excluded from the sampling frame.

resulted in the agreement of 55 FSAs to participate in the evaluation.[4] A sample of 12,000 potential FSETP participants was then randomly selected within these experimental sites and randomly assigned to the program or a control group.

As this example illustrates, one of the potential barriers to obtaining a representative sample of the population of interest is the need to obtain the cooperation of local program staff. Program staff typically resist participating in social experiments for a variety of reasons, including the added burden of experimental sample intake and random assignment procedures, fear of disruption of ongoing program activities, and ethical concerns about denial of service to controls.[5] If the refusal rate is high among selected sites, selection bias can creep into the impact estimates via self-selection of sites. For example, if only the most effective schools agree to participate in an experimental test of a remedial education program, the experimental estimates are likely to overstate the impacts of the program.

In the face of the expected cost and difficulty of recruiting a randomly selected sample of sites that is representative of the population of interest, many social experiments have opted instead for **convenience samples** of sites that, for one reason or another, are easy to recruit. Often these are sites that have expressed interest in participating in the experiment or that have established relationships with the researchers or funding agency for other reasons. In other cases, where the visibility and added resources associated with participation in a demonstration project are viewed as a benefit to the local program, sites have been selected by sponsoring agencies on political grounds. Often, such selections are a *fait accompli* before the research team has been selected.

At best, convenience samples of sites leave the experimenter with no knowledge of the relationship between the estimated program impacts in the experimental sites and what those impacts would be in the broader population of interest for policy. At worst, by concentrating the experimental sample within a self-selected set of sites, they inject the very selection bias that social experiments are intended to avoid. In most cases it is, of course, possible to compare the characteristics of the experimental sites and participant sample with those of the broader population from which they were drawn. Such comparisons can identify ways in which the experimental sample *differs* from the population of interest. But they can never demonstrate conclusively that it is truly representative of that population because it is always possible that the two differ in unmeasured characteristics that affect the outcomes of interest.

An alternative to both random selection of sites and convenience samples that is sometimes used is **purposive selection** of sites that are well-matched to the population of interest in observable characteristics. For example, sites for the Washington State Self-Employment and Enterprise Development (SEED) Demonstration were selected by choosing the combination of sites that minimized a weighted index of differences between the sites and the state overall on a number of characteristics.[6]

This approach is an improvement over convenience samples of sites in that it assures that the experimental sites are well-matched to the overall population on at least the most salient observable characteristics. Indeed, it can be argued that purposive selection is preferable to random selection of sites when the number of sites is small because in small samples sampling error can create large differences between the sample and the population from which it was drawn.[7] Purposive selection directly controls such differences in observable characteristics. And if sites are selected *solely* on the basis of observable characteristics, there is no reason to expect systematic differences in *unobservable* characteristics between the study sites and the overall population once they are matched on observable characteristics (as there is when the sites are self-selected or selected on political grounds). The principal disadvantage of purposive selection is that, unlike random selection, there is no way to quantify the sampling error involved. And as with random sampling, the experimenter still has to convince local program staff in the purposively selected sites to participate in the experiment, and any refusals to participate can inject selection bias into the sample.

A final site selection strategy that is sometimes used is purposive selection of sites that represent different social,

[4] Among the original 60 randomly selected sites, 13 refused to participate in the study, 6 of them in 3 states that refused at the outset to participate. Seven of the selected sites were found not to be implementing the program in the study year and were dropped from the sample. Backup sites were randomly selected for sites that refused. However, time constraints limited the site selection and recruiting process, and it was ultimately decided to implement the study in a sample of 55 sites. Subsequent problems with random assignment procedures in 2 sites reduced the final sample to 53 sites. See Puma et al. (1990) for further details.

[5] In a subsequent paper, we discuss ways to address these concerns.

[6] See Orr et al. (1989).

[7] The great sampling statistician Leslie Kish put the matter this way: "If a research project must be confined to a single city in the United States, I would rather use my judgment to choose a 'typical' city than select one at random. Even for a sample of 10 cities, I would rather trust my knowledge of U.S. cities than a random selection. But I would raise the question of enlarging the sample to 30 or 100 cities. For a sample of that size a probability selection should be designed and controlled with stratification." (Kish, 1965)

economic, or programmatic environments in dimensions thought to affect the impact of the program, rather than to match the distribution of those characteristics in the overall population. For example, in testing a training program for welfare recipients one might try to pick some sites with high welfare benefit levels and some with low benefits, some sites in areas with high unemployment rates and some in areas with low unemployment rates, etc. Such an approach can help researchers to understand how the impact of the experimental program varies with these conditions. But if these conditions do influence program impacts *and* their distribution among the sample sites differs from their distribution in the population of interest for policy, then the experimental estimate of the average program impact in the sample will not be an unbiased estimate of the average impact that could be expected in the broader population. To obtain an unbiased estimate of what the impact would be in the broader population, one would have to "reweight" the sample to reflect the composition of that population in these dimensions. Doing so will reduce the precision of the impact estimates relative to the estimates that would have been obtained from a more representative sample. This approach also suffers from the other shortcomings of purposive sampling discussed above.

Achieving externally valid impact estimates requires not only that the experimental sites be representative of all sites in the broader population of interest, but also that the sample of individuals within those sites be representative of that population. As discussed in the previous paper in this series, this means that the intake and random assignment process must be designed to yield a sample of the relevant population—whether that is the overall target population, eligible applicants, or potential participants. It also means that the intake process must be designed to be as similar as possible to that which would be employed in an ongoing program, or—in the case of an evaluation of an ongoing program—that the implementation of the experiment disturb the existing intake process as little as possible.

In practice, it is often extremely difficult to achieve an externally valid experimental sample. Experimenters often lack the resources needed to recruit a truly representative sample of sites and to compel or induce local program administrators in all of the selected sites to participate in the experiment. The results of their efforts in this regard must be judged not only in comparison to the ideal of a perfectly representative sample, but also in comparison to the strengths and weaknesses of the alternative available evidence. If the only alternative source of information for policymakers is the anecdotes and success stories of local program operators, experimental evidence from even a badly nonrepresentative convenience sample may be an enormous contribution. The choice would be more difficult if the alternative were a nonexperimental study based on nationally representative data on the population of interest. In that case, one would have to weigh the risks of using a potentially *internally* invalid method (the nonexperimental estimator) against the risks of using a potentially *externally* invalid method (the experiment). There is little that can be said in general about this tradeoff; each case must be examined on its own merits.

A final point that must be recognized is that the exact policy context within which the experimental results will be used is often not known when the experiment is designed. It is therefore critical that the experimental treatment and sample selection procedures be carefully documented, so that future policymakers will know how closely they correspond to the program or policy with which they are concerned and its intended target population.

Sample Size and the Statistical Power of the Design

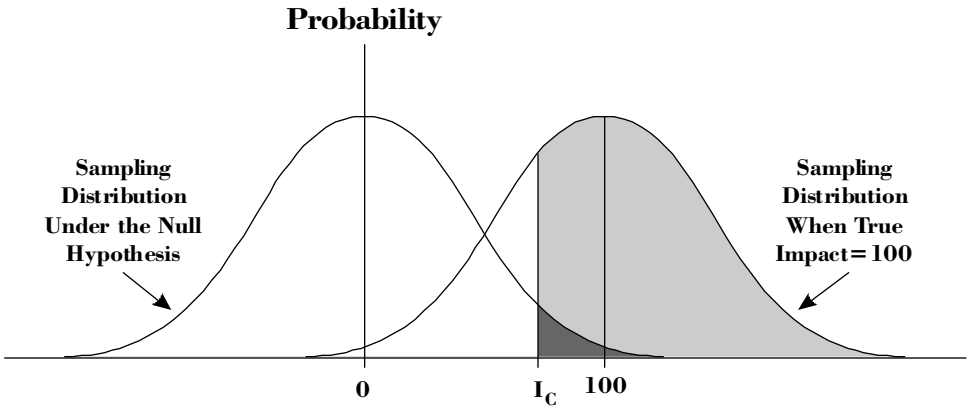
As explained in the second paper in this series, even the best designed study—either experimental or nonexperimental—cannot measure the *exact* impact of a program, or even say with certainty whether the program had an impact at all. What a well-designed experiment *can* do is to provide an unbiased estimate of the impact, say whether we can be confident that the impact is nonzero, and specify a confidence interval around the estimate within which we can be reasonably certain the true impact falls. In designing an experiment, one of our central objectives is to ensure that the confidence with which we can say whether the program had a nonzero impact is great enough, and the interval within which we can bracket the true impact is narrow enough, for policy purposes. These objectives are captured by the statistical concept of the **power of the design**.

Measuring the Power of the Design

The power of the design is *the probability that, for a specified value of the true impact, we will reject the null hypothesis of zero impact.* Suppose, for example, that we want to estimate the impact of a training program on its participants' earnings. If the true impact of the program is positive, we would like the test of statistical significance of the experimental estimate to reject the null hypothesis of zero effect. The greater the probability that it will do so, the greater is the power of the design.

Derivation of the Power of the Design, For True Impact = 100

EXHIBIT 1



Power of the Design, For True Impact = 100, With Larger Sample Size

EXHIBIT 2

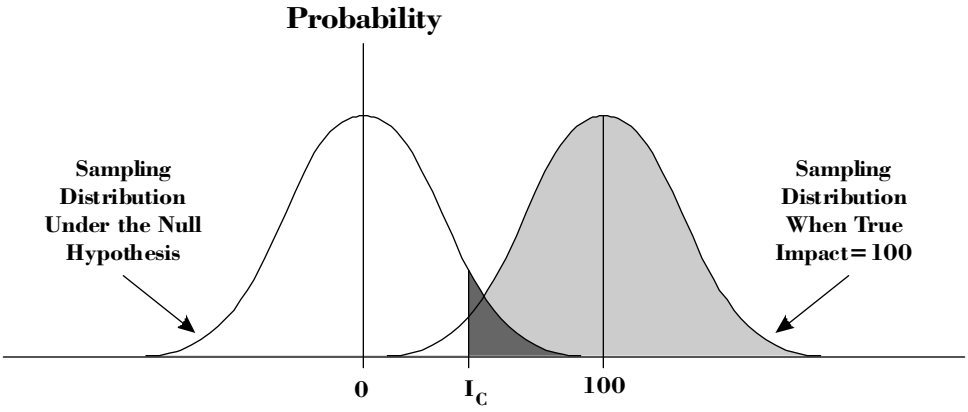


Exhibit 1 shows how the power of the design can be calculated for one specific value of the true impact, say a $100 increase in earnings. That is, it shows how to calculate the probability that, if the *true* effect of the experimental program is to increase earnings by $100, the experimental impact estimate will be significantly greater than zero. The normal curve to the left of the exhibit is the sampling distribution of the impact estimator under the null hypothesis that the true impact is zero. The dark-shaded area under the right-hand tail of that distribution is the critical region for the test of the null hypothesis at the 5 percent significance level.[8] (Note that one must specify the significance level of the test in order to calculate the power of the design.) As explained in the second paper in this series, if the experimental estimate falls in the critical region, we reject the null hypothesis of zero impact. To calculate the power of the design, then, we must determine the probability that the experimental estimate will fall in the critical region when the true impact is $100.

The answer to that question is given by the sampling distribution of the experimental estimate when the true impact is $100; this is the distribution to the right in Exhibit 1. It is centered on $100 and its shape is determined by the standard error of the experimental estimator. The probability that the experimental estimate will fall in the critical region if the true impact is $100 is given by the shaded area under this curve to the right of the critical value I_c.

[8] The exhibit develops the power of the design for a one-tailed test. Similar reasoning applies to two-tailed tests, although that case is more difficult to show graphically.

This probability is the power of the design for a true impact of $100—in this illustrative example, 70 percent. That is, there is a 70 percent chance that we would reject the null hypothesis of zero effect when the true effect is $100. (Later in this paper, we will explain how the numerical value of this probability is calculated; here, our interest is in its conceptual derivation.)

Power and Sample Size

As this example makes clear, the power of the design depends on the shape of the two sampling distributions in Exhibit 1. And as noted in the second paper in this series, the shape of the sampling distribution of the experimental estimate depends on the size of the experimental sample. In particular, the larger the experimental sample, the more tightly the sampling distribution will be clustered around its mean. Exhibit 2 shows what happens to the power of the design to detect an impact of $100 if we use a larger experimental sample than the sample in Exhibit 1. As shown in the exhibit, the tighter sampling distribution around a true impact of $100 *increases* the probability of the experimental estimate exceeding *any* value to the left of $100, including the critical value for the test of significance. Moreover, the tighter sampling distribution around the null hypothesis of zero impact *lowers* the critical value for the significance test; this also has the effect of increasing the proportion of the area under the sampling distribution around $100 that lies above the critical value (compare the shaded area under the right-hand curve in Exhibit 2 with the corresponding shaded area in Exhibit 1). For both reasons, then, increasing the size of the experimental sample increases the power of the design.[9]

Power and the Significance Level of the Test

A second way in which we could increase the power of the design would be to raise the significance level for the test of the null hypothesis of no effect. Suppose, for example, that instead of testing at the 5 percent significance level, we were to test at the 10 percent significance level. This would lower the critical value, I_C, thereby increasing the proportion of the area under the right-hand sampling distribution that falls in the critical region—*i.e.*, it would increase the probability of rejecting the null hypothesis of zero effect when the true effect is $100.

Note, however, that raising the significance level of the test also *increases the probability of rejecting the null hypothesis when it is in fact true* from 5 percent to 10 percent. Thus, in specifying the significance level of the test, there is a tradeoff between two risks: the risk of falsely concluding that there is a positive effect (*i.e.*, rejecting the null hypothesis) when in fact there is no effect and the risk of failing to reject the null hypothesis of zero effect when in fact the true effect is positive. The probability of the former error is given by the significance level of the test of the null hypothesis. The probability of the latter error is one minus the power of the design.[10] Thus, in the design depicted in Exhibit 1, we run a 5 percent risk of falsely concluding that the program effect was positive when in fact it was zero and a 30 percent risk of falsely concluding that the program effect was zero when it was in fact $100.

In making the tradeoff between these two risks, researchers typically accept a higher risk of mistakenly concluding that the effect is zero than of falsely concluding that it is positive, on the grounds that the costs of the latter error are greater than the costs of the former error. Suppose, for example, we are testing a new program which, if found to be effective, will be implemented on an ongoing basis, at a cost of $100 million per year. If we mistakenly conclude that the program is effective when in fact it has zero effect, over time billions of dollars will be wasted on it. If we make the converse error—concluding that the program has zero impact when its true effects are positive—we miss an opportunity to implement an effective program, but we do not waste large sums of money.[11]

However one views these risks, it is essential that the tradeoff be made explicitly. Far too often, researchers unthinkingly apply the "conventional" significance levels of 5 or 10 percent without examining the implications for the power of the design. The result can be an extremely weak test of the null hypothesis—*i.e.*, only a low probability of detecting a positive impact if it exists. In such cases, one should consider increasing the sample size to strengthen the design, lowering the significance level to achieve a better balance between the two types of risk, or—if it is not possible to obtain a sufficiently large sample to yield adequate power—not conducting the experiment at all.

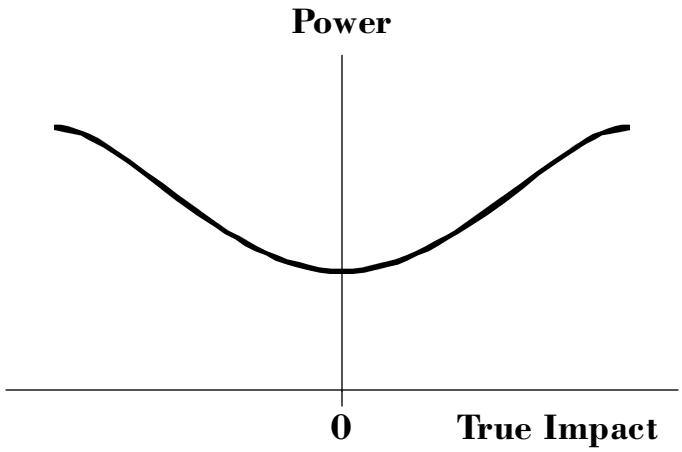
[9] When the critical value is initially above the mean of the right-hand distribution in Exhibit 1, these two effects are offsetting. It can be shown, however, that their net effect is to increase the power of the design. Thus, an increase in sample size *always* increases the power of the design.

[10] Rejecting the null hypothesis when it is true (i.e., falsely concluding that there is a positive effect) is known as a Type-I error. Failing to reject the null hypothesis when it is false (i.e., failing to detect a true positive effect) is known as a Type-II error.

[11] Similar reasoning applies to the evaluation of an ongoing program. A false positive results in the continuation of a waste of resources, whereas a false negative results in unnecessarily terminating the program. The social cost of the latter will depend on the alternative use of the resources formerly devoted to the program.

Power Function

EXHIBIT 3



Power Functions

The discussion so far has been cast in terms of the power of the design at a *single* positive value of true impact. Obviously, the power of the design can be calculated similarly for any value of true impact. If power is calculated for all possible levels of true impact, the resulting probabilities trace out the **power function** for the design. Exhibit 3 shows an illustrative power function. The height of the curve measures the power of the design for each value of true impact (the horizontal axis). The curve has the characteristic shape of power functions, with a minimum when true impact is zero (where the null hypothesis is true) and power rising asymptotically toward 1.0 at values of true impact further away from zero.[12]

The power function is conditional on the significance level of the test and the sample size. For a given significance level, the power function corresponding to a larger sample size will lie above the power function for a smaller sample size—*i.e.,* the larger sample size will have greater power for all true impact values (except zero). For a given sample size, the power function for a higher significance level (say, 10 percent) will lie entirely above the power function for a lower significance level (say, 5 percent); this illustrates the tradeoff between the two types of risk discussed above.

Minimum Detectable Effects and the Design of Experiments

Choosing the sample size and significance level for an experiment is equivalent to choosing the power function for the experiment. In practice, however, rather than attempting to calculate all the possible power functions corresponding to different sample sizes and significance levels, experimenters typically specify the *desired* power for a *specified* value of true effect and significance level of the test, then solve for the sample size that will yield that level of power.[13] To do so, they use the concept of **minimum detectable effect**—the smallest *true* impact that would be found to be statistically significantly different from zero at a specified level of significance with specified power.

Suppose that in the case of the training program discussed above, we would like to have an 80 percent probability of detecting a true impact of $100 if it occurs and that we are willing to take a 10 percent risk of rejecting the null hypothesis of zero when it is in fact true. We therefore want to know the sample size that will yield a minimum detectable effect of $100 with 80 percent power at the 10 percent significance level. Exhibit 4 shows how we can calculate this sample size.[14]

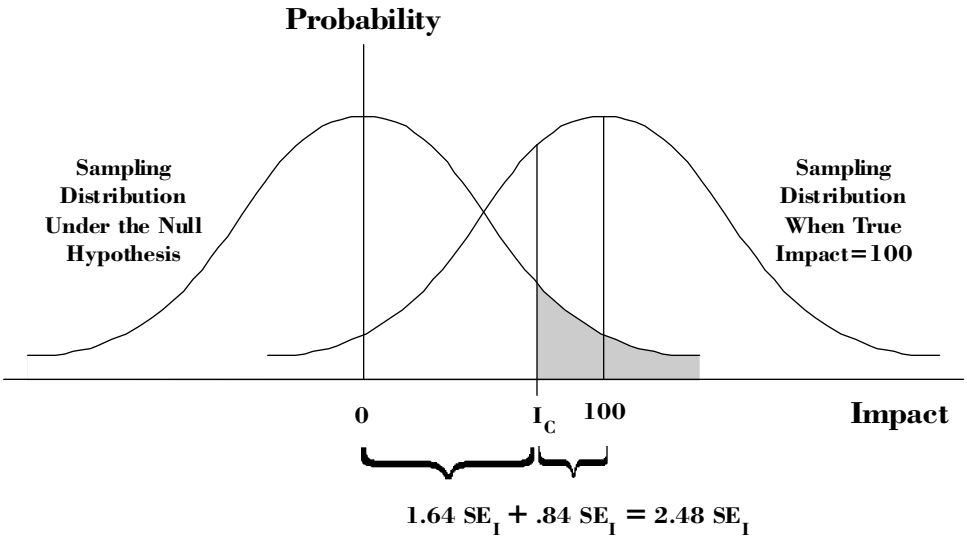
[12] The minimum value of the power function is equal to the significance level and occurs at zero. This follows from the definition of power as the probability of rejecting the null hypothesis when the true effect is I_0. When I_0 is zero, then this probability is the same as the probability of rejecting the null hypothesis of zero effect when it is true--i.e., the significance level.

[13] This is equivalent to specifying the family of power functions defined by a given significance level and then choosing among them on the basis of their levels at a single value of true impact.

[14] This method of illustrating the derivation of minimum detectable effects is adapted from Bloom (1995).

Calculation of Minimum Detectable Effects

EXHIBIT 4



As before, the left-hand curve in the exhibit is the sampling distribution of the experimental estimate under the null hypothesis of zero effect; the shaded region in its right tail is the critical region for rejection of the null hypothesis at the 10 percent significance level. In large samples, the critical value that defines this region (I_C will be 1.64 times the standard error of the impact estimate (SE_I). The right-hand curve in the exhibit is the sampling distribution of the experimental estimate when the true impact is $100. To have an 80 percent probability of rejecting the null hypothesis of zero effect when the true impact is $100, 80 percent of the area under this sampling distribution must lie to the right of the critical value 1.64 SE_I. This will be the case when the mean of the distribution ($100) lies 0.84 standard errors above the critical value.[15] As shown in the exhibit, then, to make this condition hold exactly we need only choose the sample for which $100 equals 2.48 (=1.64+.84) times the standard error of the impact estimate. That is, we choose n_T and n_C, the numbers assigned to the treatment and control groups, so that:

1
$$100 = 2.48\,SE_I = 2.48 \cdot \sqrt{\frac{V_Y}{n_T} + \frac{V_Y}{n_C}}$$

where V_Y is the variance of Y, the outcome of interest.

Alternatively, for any specified combination of treatment and control group samples we can calculate the minimum detectable effect (MDE) achievable with 80 percent power at the 10 percent significance level:

2
$$MDE = 2.48 \cdot \sqrt{\frac{V_Y}{n_T} + \frac{V_Y}{n_C}}$$

Choice of a different level of power or significance simply changes the multiplicative constant in equations 1 and 2. For example, for 80 percent power at the 5 percent significance level, the constant would be 2.80. Thus, more generally:

3
$$MDE = k \cdot \sqrt{\frac{V_Y}{n_T} + \frac{V_Y}{n_C}}$$

where k is a constant that reflects the chosen levels of power and significance.

As is clear from equation 3, all that is required to compute the minimum detectable effect for any given value of k and combination of treatment and control sample sizes is knowledge of the variance of the outcome. This is usually available from existing data. For example, the variance of the earnings of low-income workers can be computed on the basis of data from nonexperimental evaluations of training programs or from national surveys like the Current Population

[15] The numerical values in this example are obtained from a standard table of values of the t-statistic. Recall that the t-statistic is defined as the impact estimate divided by its standard error.

Survey.[16] In using existing data sources for this purpose, it is of course important to ensure that the population represented in the data is closely similar to the planned experimental population.

An important property of experimental designs is readily derived from equation 3: *For any given division of the sample between treatment and control groups, minimum detectable effects are inversely proportional to the square root of the overall sample size*.[17] To see this, let s_T be the share of the total sample N allocated to the treatment group and s_C be the share of the total sample allocated to the control group. Then equation 3 can be written as:

4

$$\text{MDE} = k \cdot \sqrt{\frac{V_Y}{s_T N} + \frac{V_Y}{s_C N}}$$

$$= k \cdot \left(\sqrt{\frac{1}{N}}\right) \cdot \sqrt{\frac{V_Y}{s_T N} + \frac{V_Y}{s_C N}}$$

Thus, for example, doubling the sample reduces the minimum detectable effect by a factor of $1 \div 1.41 = 0.71$. To halve the minimum detectable effect, one would have to quadruple the sample.

The minimum detectable effect is an extremely useful indicator of the power of any particular design. Small minimum detectable effects mean that policy makers can be quite confident that if the program has even a small effect on the outcome, the experiment will have a good chance of detecting it (*i.e.*, of rejecting the null hypothesis of no effect). Large minimum detectable effects mean that the effect of the program would have to be large for the experiment to have a good chance of detecting them. Prior analysis of the power of the design is the best protection against ending up in the situation described in an earlier paper in this series—obtaining experimental estimates with confidence intervals so broad that they are consistent with both large effects and no effect at all.

Of course, "small" and "large" are relative terms, and there is no obvious way to decide what size effects we want to be reasonably sure of detecting. One rule that is sometimes suggested is to set the minimum detectable effect at the level that would make the program cost-effective—*i.e.*, design the experiment so that if the program is cost-effective we can be reasonably certain that the experiment will find a nonzero effect.[18] While this rule has a certain intuitive appeal, it ignores the relationship between minimum detectable effects and the cost of the experiment. Achieving smaller minimum detectable effects requires larger samples, which increases the cost of the experiment. Thus, a truly general rule for deciding on sample sizes and the power of the design would have to specify the tradeoff between experimental costs and the social value of more powerful estimates of program impact.[19] In the absence of such a general rule, the best advice that can be given is that the experimenters and policymakers compute and review a "menu" of alternative designs, with different costs and minimum detectable effects on the outcomes of central interest, in order to make a judgmental tradeoff between power and cost.

The Point of Random Assignment and the Power of the Design

In some cases the power of the design will be influenced by the design of the random assignment process, as well as by sample size. This will be the case when interest focuses on estimating program impacts on participants and not all of those randomly assigned participate in the program. In this section, we discuss this case.

As shown in the second paper in this series, under the assumption that the program had no effect on nonparticipants, an unbiased estimate of program impact on participants (I_P) can be obtained by dividing the estimated impact on the overall treatment group (I_T) by the participation rate among those randomly assigned (r):

5

$$I_P = \frac{I_T}{r}$$

[16] For one important class of outcomes, the variance can be computed from the mean of the outcome. For dichotomous outcomes (i.e., outcomes that can take on only two values, 0 or 1), the variance of the outcome is *m(1-m)*, where *m* is the sample mean. (Since *m* is also the proportion of the sample taking on the value 1, it can also be thought of as the proportion of the sample with positive outcomes.) The expression *m(1-m)* is maximized when *m*=.5. Therefore, even if the mean is unknown, one can compute the worst case minimum detectable effect for such an outcome. Outcomes like "completed high school", "employed at follow-up", and "left welfare" belong to this class.

[17] Later in this paper we discuss the optimal allocation of the sample between the treatment and control groups.

[18] This assumes that program costs can be predicted, which is often the case, at least approximately.

[19] A framework within which this could be done is presented in Burtless and Orr (1986).

The standard error of the estimated impact on participants (SE_{IP}) can be derived from the standard error of the estimated impact on the overall treatment group (SE_T) by the same procedure:[20]

$$\textbf{6} \qquad SE_{IP} = \frac{SE_{IT}}{r}$$

Our measure of the power of the design, the minimum detectable effect on participants (MDE_{IP}), is:

$$\textbf{7} \qquad MDE_{IP} = k\, SE_{IP} = k \frac{SE_T}{r}$$

Thus, the minimum detectable effect on participants depends not only on sample size and the variance of the outcome (which determine SE_T), but also on the participation rate among those randomly assigned. The lower the participation rate, the larger will be the minimum detectable effect on participants—*i.e.*, low participation rates result in low statistical power.

The experimenter can influence the participation rate, and thereby improve the power of the design, through the design of the random assignment process. Exhibit 5 shows the intake process for a voluntary program.[21] As can be seen in the exhibit, individuals drop out of the intake process at various points between the initial response to outreach and participation in the program. Since the participation rate r in equations 5-7 is defined as the proportion *of those randomly assigned* who participate in the program, this means that the later in the intake process random assignment is administered, the higher the participation rate will be. Thus, *the principal way the experimenter can increase the participation rate and reduce minimum detectable effects is by conducting random assignment as late in the intake process as possible.*[22]

Consider, for example, the choice between randomly assigning all applicants and randomly assigning only *eligible* applicants. Suppose that the ineligibility rate is 20 percent, so that for every 100 applicants there are only 80 eligible applicants, and that of those 80, 60 would ultimately participate in the program in the absence of random assignment. The participation rate for applicants, then, is 0.60 (= 60÷100), whereas the participation rate for eligible applicants is 0.75 (= 60÷80). As shown in equation 7, the minimum detectable effect on participants for a sample composed of applicants is:

$$\textbf{8} \qquad MDE_{IP,A} = k \frac{SE_T}{r_A} = k \frac{SE_T}{.60}$$

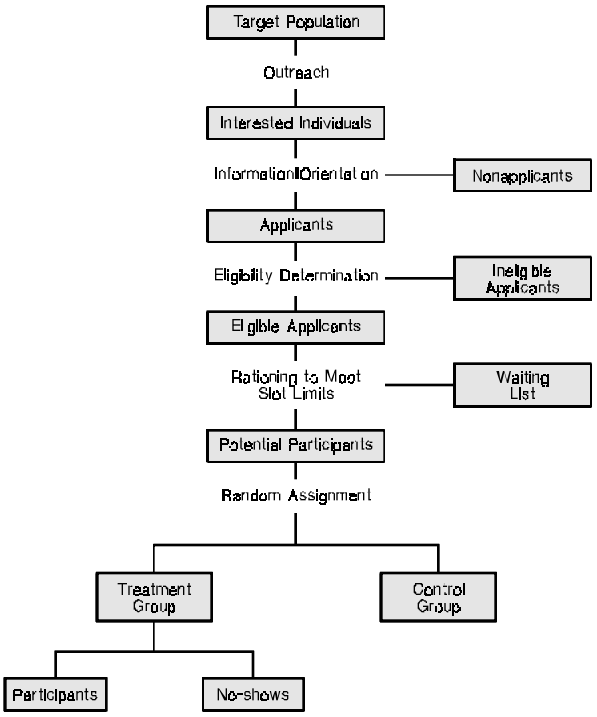
and the minimum detectable effect on participants for a sample composed of eligible applicants is:

$$\textbf{9} \qquad MDE_{IP,E} = k \frac{SE_T}{r_E} = k \frac{SE_T}{.75}$$

where r_A and r_E are the participation rates of applicants and eligible applicants, respectively.

Design for Estimating Impacts on Participants in a Voluntary Program

EXHIBIT 5



[20] This assumes that the participation rate would be constant across replications of the experiment. This is almost certainly not strictly true, but if each individual has a constant *probability* of participating, in large samples the variation of the overall participation rate will be so small as to be negligible.

[21] This exhibit is adapted from Exhibit 1 in the third paper in this series.

[22] It might appear that the participation rate could also be increased by taking administrative measures to reduce the number of individuals dropping out of the intake process—*e.g.*, by following up with individuals who fail to apply for the program and encouraging them to do so, or by tracking down no-shows and encouraging them to participate. However, if such steps would not be taken in an ongoing program, taking them in the experiment would result in a different composition of the participant population from that which would occur in an ongoing program, thereby undermining the external validity of the experiment.

The relative power of the two designs is indicated by the ratio of these two minimum detectable effects:

$$\textbf{10} \qquad \mathrm{MDE}_{\mathrm{IP,E}} = \frac{k\left(\frac{\mathrm{SE}_{\mathrm{T}}}{\mathrm{r}_{\mathrm{A}}}\right)}{k\left(\frac{\mathrm{SE}_{\mathrm{T}}}{\mathrm{r}_{\mathrm{A}}}\right)}$$

$$= \frac{\mathrm{r}_{\mathrm{E}}}{\mathrm{r}_{\mathrm{A}}}$$

$$= \frac{.75}{.60} = 1.25$$

Thus, random assignment of all applicants results in minimum detectable effects on participants that are 25 percent larger than if only eligible applicants are randomly assigned.

Similarly, random assignment of potential participants will result in smaller minimum detectable effects on participants than random assignment of eligible applicants. The relative power of these two designs can be calculated in the same fashion as the relative power of assigning all applicants and assigning only eligible applicants.

It might appear that the loss in power associated with assigning all applicants could be offset simply by randomly assigning 25 percent more applicants, so that the number of participants in the program is the same under the two approaches. In fact, the increase in sample size would have to be much larger than 25 percent. To see this, first note that for these two designs to have the same power, one would have to set their relative sample sizes so that:

$$\textbf{11} \qquad \mathrm{MDE}_{\mathrm{IP,A}} = k\left(\frac{\mathrm{SE}_{\mathrm{T,A}}}{\mathrm{r}_{\mathrm{A}}}\right)$$

$$= \mathrm{MDE}_{\mathrm{IP,E}} = k\left(\frac{\mathrm{SE}_{\mathrm{T,E}}}{\mathrm{r}_{\mathrm{A}}}\right)$$

which implies:

$$\textbf{12} \qquad \frac{\mathrm{SE}_{\mathrm{T,A}}}{\mathrm{SE}_{\mathrm{T,E}}} = \frac{\mathrm{r}_{\mathrm{E}}}{\mathrm{r}_{\mathrm{A}}}$$

$$= .80$$

That is, to achieve the same power, the sample of all applicants would have to be sufficiently larger to reduce the standard error of the impact estimate by 20 percent. But since the *variance* of the impact estimate is inversely proportional to sample size, the *standard error* of the impact estimate is inversely proportional to *the square root of the sample size*. This means that the applicant sample would have to be 56 percent ($= 1 \div .80^2 - 1$) larger than the eligible applicant sample to achieve the same power.

While the numbers used in this illustrative example are purely hypothetical, they are typical of the orders of magnitude involved in the choice of placement of random assignment in the intake process. As these numbers suggest, this choice can have a substantial effect on the power of the design or, if sample size is increased to offset the loss of power, the cost of the experiment. In the example, assigning eligible applicants rather than all applicants would entail either a 25 percent increase in minimum detectable effects or a 56 percent increase in sample in order to maintain the same power. In most cases, an increase in sample size of this magnitude would increase the costs of implementing the experiment and collecting data by nearly the same factor. In any specific experiment where impacts on participants are to be estimated, then, it will be important to predict the likely participation rate, so that realistic estimates of minimum detectable effects can be derived, and to consider how participation rates and minimum detectable effects will vary with the placement of the point of random assignment.

Unfortunately, program staff typically resist placing random assignment late in the intake process. Late random assignment increases the burden on staff, as they must continue to process those who will ultimately be assigned to the control group, in addition to those who will be allowed to participate in the program. Moreover, staff often feel that it is unfair to applicants to require them to invest additional time and effort, and to raise their expectations, only to be assigned to the control group. Finally, intake staff find it much more difficult to inform controls that they are excluded from the program after they have had extensive contact with them. For these reasons, it may ultimately be necessary to conduct random assignment at a point in the intake process that is not absolutely the latest point at which it could occur. But experimenters must be cognizant of the analytic losses and/or monetary costs involved in such a compromise decision.

Allocation of the Sample Among Multiple Treatments

The essence of social experimentation is comparison of outcomes among randomly assigned groups of individuals. In this section, we discuss how the relative number of individuals to be randomly assigned to each experimental group is determined. We begin with the simple case of a single treatment group and a single control group, then generalize the analysis to multiple treatment groups. We next consider the allocation of the sample when experimental costs vary from one experimental group to another. Finally, we discuss the issues that arise when multiple treatments are implemented in multiple sites.

Allocating the Sample Among Experimental Groups

The objective of sample allocation is to maximize the power of the design. Thus, in a simple experiment with one treatment group and one control group, we want to choose n_t and n_c to yield the smallest possible minimum detectable effect. In a previous section we showed that, for any given allocation of the sample between the treatment and control groups, the minimum detectable effect is inversely proportional to total sample size. Here, we hold total sample size constant and focus on the *allocation* of the sample among experimental groups; *i.e.*, we pose the problem in terms of choosing the *ratio* n_T/n_C for any given total sample. Thus, we wish to choose n_T/n_C to minimize:

13 $$\text{MDE} = k \cdot \sqrt{\frac{V_Y}{n_T} + \frac{V_Y}{n_C}}$$

It can be shown that this expression is minimized when $n_T/n_C = 1$, *i.e.*, when equal numbers of individuals are assigned to the treatment and control groups.

The way we allocate the sample among experimental groups is, of course, through random assignment. The desired sample allocation determines the **random assignment ratio**—the ratio of the probability that a given individual will be assigned to one group to the probability that he or she will be assigned to another. In the simple two-group case just described, the optimal treatment-control random assignment ratio is 50/50—*i.e.*, a 50 percent chance of being assigned to each group. Random assignment ratios need not be equal across experimental groups. As we will see, in some cases the optimal sample allocation assigns very different numbers of individuals to different experimental groups; in those cases, random assignment probabilities would vary commensurately.

When there are multiple treatment groups, we face a tradeoff among experimental objectives. Within a fixed total sample, allocating more of the sample to one treatment group will increase the power of the design for estimating the impact of that treatment *at the expense of the power of the design for other treatments*. To determine the optimum allocation in this situation, we must specify the importance we attach to each of the impact estimates to be derived. We do this by specifying an objective function W that is a weighted sum of the minimum detectable effects for the k treatments:

14 $$\text{MDE} = w_1\text{MDE}_1 + w_2\text{MDE}_2 + \ldots + w_k\text{MDE}_k$$

where w_I is the "policy weight" attached to the impact estimate for the Ith treatment. Since smaller minimum detectable effects are preferred to larger ones, we wish to allocate the sample to *minimize* W, subject to the constraint:

$$n_1 + n_2 + \ldots + n_{k+1} \leq N$$

where n_I is the number of individuals assigned to the Ith of k+1 experimental groups (k treatment groups, plus a control group) and N is the total sample size.

Consider, for example, an experiment with two treatment groups and a control group. Such an experiment can produce two different types of impact estimate: the impact of treatment 1 and the impact of treatment 2. If we put equal weight on these two different estimates (*i.e.*, if $w_1 = w_2 = 1$), then we wish to minimize:

15 $$\begin{aligned} W &= \text{MDE}_1 + \text{MDE}_2 \\ &= \sqrt{\frac{V_Y}{n_{T1}} + \frac{V_Y}{n_C}} + \sqrt{\frac{V_Y}{n_{T2}} + \frac{V_Y}{n_C}} \end{aligned}$$

subject to the constraint:

$$n_{T1} + n_{T2} + n_C \leq N$$

It can be shown that W is minimized (*i.e.*, the power of the design is maximized) when $n_{T1}/n_{T2} = 1$ and $n_C/n_{T1} = n_C/n_{T2} = 2$. That is, in the optimal allocation the samples assigned to the two treatment groups are of equal size and the sample assigned to the control group is *twice* as large as each of the treatment groups. This means that half the sample should be assigned to the control group and one quarter to each treatment group.

To see why we obtain this asymmetric result, consider the effect of adding one individual to each of the three experimental groups. Adding one individual to treatment group

1 reduces the minimum detectable effect for that treatment (the first term in equation 15), but has no effect on the minimum detectable effect for treatment 2 (the second term); the converse holds for adding an individual to the sample assigned to the second treatment. In contrast, adding an individual to the control group reduces the minimum detectable effect for *both* treatments, because the control group is involved in both experimental comparisons. Of course, since sample size enters into the *denominator*, the larger the sample already assigned to a particular group, the less difference the addition of one more individual will make. If one thinks of starting with an equal allocation among the three groups and then shifting sample from the two treatment groups to the control group, it turns out that the reduction in minimum detectable effect resulting from each additional control group member exceeds the increase in minimum detectable effect resulting from the loss of a treatment group member until the control group is exactly twice the size of each of the treatment groups.

This result for the three-group case can be generalized to a simple rule that applies to any number of groups as long as equal weights are placed on each experimental comparison: *allocate the sample in proportion to the number of experimental comparisons in which each group is involved*. If, for example, the impact of seven different experimental programs are to be derived on the basis of seven treatment groups and one control group, each treatment group will be involved in one experimental comparison and the control group will be involved in seven comparisons. Thus, if there is equal policy interest in each of these comparisons, the optimal allocation would place 1/14 of the sample in each treatment group and one half (7/14) of the sample in the control group.

If there is *unequal* policy interest in the different experimental comparisons, the sample allocation must be derived by minimizing the expression for W in equation 15, subject to the constraint that the sum of the samples assigned to the various experimental groups cannot exceed the fixed total sample.

Sample Allocation Subject to a Fixed Budget

Up to this point, we have taken a fixed total sample to be the constraint on sample allocation. This will sometimes be the case, as in an experimental evaluation of an ongoing program in which all eligible applicants who apply within a given time period are to be randomly assigned. More commonly, however, the binding constraint is not a fixed total sample size, but a fixed budget that can be devoted to the experimental treatments and data collection. In that case, minimum detectable effects should be minimized subject to the constraint of a fixed budget, rather than a fixed total sample size.

If the cost of assigning an individual to one experimental group is the same as that of assigning him or her to any other, then having a fixed budget is the same as having a fixed sample size; the total sample size is simply the budget divided by the (uniform) cost of assigning an individual to an experimental group. A common situation in which this is the case is that where the experimental treatments are not funded through the budget of the agency sponsoring the experiment. This would be the case, for example, for evaluations of ongoing programs, where the treatment is provided out of the regular program budget and the evaluation is funded through a separate research budget. In that case, the cost *to the evaluation budget* of assigning an individual to an experimental group is simply the cost of collecting data on that individual; since the same data are collected for all experimental groups, this cost does not vary from one group to another.

If the cost of treatment is included in the experimental budget, then the cost per sample member will generally vary from one experimental group to another. When multiple treatments are tested, some are likely to cost more than others; in any case, costs per sample member are likely to be higher in the treatment group than in the control group, since controls receive no experimental services. Within a fixed budget, unequal costs per sample member mean that a larger sample can be supported by assigning more individuals to the cheaper experimental groups.

The general solution for the optimal sample allocation when costs vary among experimental groups is to minimize W (as defined in equation 14) subject to the constraint:

$$\mathbf{16} \qquad n_1 c_1 + n_2 c_2 + \ldots + n_{k+1} c_{k+1} \leq C$$

where n_I is the number of individuals assigned to the *I*th group, c_I is the cost per sample member in the *I*th group, and C is the total budget for experimental treatment and data collection.

While the solution to this problem is mathematically straightforward, when there are multiple treatment groups the solution is somewhat complicated. In the simple case of a single treatment group and a single control group, the optimal allocation is:

$$\mathbf{17} \qquad \frac{n_T}{n_C} = \sqrt{\frac{c_C}{c_T}}$$

That is, the sample should be allocated between the treatment and control groups in inverse proportion to the square root of the relative costs per sample member in the two groups.

A simple example will illustrate the importance of taking variations in the cost per sample member into account in sample allocation. Suppose we have a budget of $500,000 to evaluate a social service program for low-income families. For each family assigned to the treatment group, we will incur a cost of $4,000 for experimental services. Data collection will cost $500 per family regardless of experimental assignment. Thus, the total cost per treatment group family will be $4,500 and the total cost per control family will be $500. If we allocate equal numbers of families to the treatment and control groups, our budget will support 100 families in each group. Taking the relative costs of the two groups into account, however, the optimal allocation is to put three times as many families in the control group as in the treatment group. With this allocation, our budget will support 83 families in the treatment group and 250 families in the control group. Minimum detectable effects under this allocation will be 10 percent smaller than they would have been with equal-sized groups.[23] To achieve this reduction in minimum detectable effects with equal-sized groups would have required a 23 percent increase in total sample size.[24] Thus, moving from an equal allocation to the optimal allocation is equivalent to increasing the budget for the experiment by 23 percent, or $115,000.

"Unbalanced" Designs

The sample allocation procedures discussed in the preceding section take into account differences in average cost per sample member among experimental groups. In some cases, experimental costs may vary systematically not only with the experimental group to which an individual is assigned, but also with the individual's characteristics. For example, negative income tax plans provide larger payments to low-income families than to higher-income families.

If one extends the logic of the preceding section to include this source of variation in cost, in a negative income tax experiment one would assign relatively more low-income families to the treatment groups, where they are cheaper observations than higher-income families, than to the control group, where they cost the same as higher-income families. This is in fact what was done in the early income maintenance experiments.

While taking the variation of experimental cost with family type into account in the sample allocation arguably improves the power of the design that can be supported with a fixed budget, it is important to recognize that it fundamentally changes the nature of the experimental sample. In all of the allocations considered up to this point, although sample size might vary across experimental groups, the *composition* of the groups did not vary systematically—except for sampling error, each group was well-matched to every other group. When we allow assignment to experimental group to be affected by *individual characteristics*, this is no longer true—the composition of the experimental groups differs systematically *by design*. Such designs are called **unbalanced designs**, and are somewhat controversial in the literature on social experimentation.[25] In part for this reason, and in part because the cost of experimental treatments seldom varies so systematically with family characteristics, the only major social experiments to employ unbalanced designs were the early income maintenance experiments. The issues involved in the analysis of data from unbalanced designs are beyond the scope of these papers. Therefore, while we note their existence, we will continue to focus on designs in which all experimental groups are well-matched.

Allocation to Multiple Treatments in Multiple Sites

Because program impacts may vary across sites, when multiple treatments are to be tested in multiple sites it is essential to avoid confounding site effects with program effects. To take an extreme case, suppose we were to randomly assign unemployed workers in City A to classroom training or a control group, and unemployed workers in City B to on-the-job training or a control group. Comparison of the treatment and control groups in City A will provide unbiased estimates of the impact of classroom training *in City A*; likewise, comparison of the treatment and control groups in City B will yield unbiased estimates of the impacts of on-the-job training *in City B*. But if we find that, say, classroom training in City A was more effective in raising workers' earnings than on-the-job training in City B, we will not know whether to conclude that classroom training is a more effective training strategy than on-the-job training or that, because of the nature of the workers or the local economy in the two cities, it is simply easier to raise the earnings of unemployed workers in City A. Since

[23] This result is obtained by calculating the ratio of the minimum detectable effects under the two allocations, using equation 13.

[24] This result is derived from the fact that, for any given sample allocation, minimum detectable effects are inversely proportional to the square root of total sample size.

[25] For an excellent discussion of the advantages and disadvantages of unbalanced designs, see Hausman and Wise (1985).

we didn't try classroom training in City B or on-the-job training in City A, we can never distinguish between these two potential explanations. In this example, treatment and site are completely confounded.

To avoid confounding treatment and site, we would like the distribution of each experimental group across sites to be the same as that of every other experimental group. This can be achieved by randomly assigning individuals to all experimental groups, in the same proportions, in every site. If this is done, the overall samples assigned to the different treatments, and to the control group, will be well-matched and fully comparable differential impact estimates can be derived.

Perhaps the most common reason for confounding of treatment and site is that tests of different treatments are conceived and executed as independent studies—often by the same funding agency. In that situation, it is almost unavoidable that the experiments will be conducted in different sites, resulting in complete confounding of treatment and site. Unfortunately, in the policy process the results of these independent tests are nearly always treated as if they reflected only the differential impacts of the different treatments. In these cases, much more reliable guidance for policy could have been obtained at the same cost by combining the tests in a common set of sites.

In some cases, however, confounding of treatment and site is unavoidable for practical reasons. This is especially likely to be the case when a large number of treatments are being tested and the experiment is to be run through existing program agencies. Suppose, for example, that we wish to test six different approaches to increasing the employment of AFDC recipients. If all six approaches were to be implemented in the same welfare office, the staff of that office would not only have to become knowledgeable about all six treatments, but would have to consistently apply the rules of each approach to those recipients, and only those recipients, assigned to that treatment. The burden on staff and the potential for contamination of the treatment in this situation is probably untenable.

Even in this situation, however, one can avoid *complete* confounding of treatment and site if program staff are willing to administer more than one treatment in the same office. Exhibit 6 shows how the experimental sample can be allocated across sites to allow differential impacts to be estimated with no more than three of the six treatments implemented in any one site (as indicated by the X's in the exhibit). For example, as shown in the shaded area of the exhibit, the impacts of treatments 1 and 2 could be compared for well-matched samples in sites B and C. Similarly, treatments 2 and 3 could be compared in sites C and D. As can be seen from the exhibit, each treatment can be

Sample Design for Multiple Treatments in Multiple Sites, When Not All Treatments Can Be Implemented in Each Site

EXHIBIT 5

SITE	TREATMENT					
	1	2	3	4	5	6
A	X					
B	X	X				
C	X	X	X			
D		X	X	X		
E			X	X	X	
F				X	X	X

compared with all but one of the other five treatments in at least one site. Each comparison is based on only one-third or two-thirds of the sample assigned to the treatments being compared, however, depending on the number of sites involved. This is therefore a less powerful design than one in which individuals are assigned to all treatments in all sites.

Moreover, since different pairs of treatments are compared in different combinations of sites, the experimental comparisons cannot, by themselves, establish the relative magnitudes of impacts across all six treatments, or even rank order the six treatments by impact, holding site effects constant. To do so requires some assumption about the interaction of treatment and site—*e.g.*, that, although treatment effects may vary across sites, site effects do not vary across treatment. Data from this design are most conveniently analyzed by multivariate methods, which will be discussed in a subsequent paper.

The Number of Experimental Sites

In designing an experiment, one generally has a choice between concentrating the sample in a small number of sites or spreading it over a larger number of sites. If cost were not an issue, it is clear that a larger number of sites is preferable. Just as increasing the number of individuals in the sample reduces the standard error of the impact estimate by "averaging out" sampling error, including a large number of sites should average out site-specific variations in the impact of the experimental program.

But there are usually substantial costs associated with adding experimental sites. If the treatment is to be administered by existing program agencies, there are costs of recruiting those agencies and training them in the experimental procedures. If it is to be administered by the researchers themselves, it may be necessary to set up an office in each site. Similarly, if data are to be collected from sample members in person, it will be necessary to hire and train a data collection staff in each site and it may be necessary to establish a field office in each site. Even if data are to be collected from administrative systems, such as welfare records, additional sites may increase the number of systems that must be accessed.

The choice of the number of experimental sites is a tradeoff, then, between the increased power that additional sites provide and the increased costs they entail. Within a fixed budget, this is a tradeoff between sample size and number of sites, because the costs of recruiting and administering additional sites must be deducted directly from the funds available for the experimental treatment and/or data collection.

A relatively straight-forward technique is available for estimating the optimal number of sites in any particular experiment.[26] Unfortunately, it requires data that may not be available in all cases. This approach poses the problem as one of choosing the number of experimental sites (q) that minimizes the standard error of the impact estimate subject to a budget constraint that takes fixed site costs into account. If the sample is evenly distributed across sites, with n sample members in each site, the standard error of the overall impact estimate can be expressed as:

$$\textbf{18} \qquad SE_I = \sqrt{\frac{4V_Y}{nq} + \frac{V_{IS}}{q}}$$

where V_Y is the variance of the outcome of interest in the population and $4V_Y/n$ is the variance of the impact estimate within a single site.[27] V_{IS} is the variance of the experimental impact across sites (which is independent of the within-site sample size). Thus, the standard error of the estimate can be partitioned into a within-site component and a between-site component. We wish to choose q so as to minimize SE_I subject to the following budget constraint:

$$\textbf{19} \qquad C = q\,(C_S + cn)$$

where C is the total budget for the experimental treatment and data collection, C_S is the fixed cost of each additional site, and c is the marginal cost per sample member.

This formulation makes clear the tradeoff between sample size and number of sites within a fixed budget. Other things equal, increases in the number of sites (q) reduce the standard error of the impact estimate, as shown in equation 18. But, as shown in equation 19, increasing q also increases the fixed site costs of the experiment (qC_s), so that to stay within a fixed budget, total sample size (qn) must be reduced.[28] This increases the within-site variance com-

[26] This technique is due to Morris (1974).

[27] This follows directly from the following expression for the variance of the within-site impact estimate (V_{IW}), setting $n_t = n_c = n/2$:

$$\begin{aligned} V_{IW} &= [(V_Y / n_t) + (V_Y / n_c)]^{1/2} \\ &= [(2V_Y / n) + (2V_Y / n)]^{1/2} \\ &= (4V_Y / n)^{1/2} \end{aligned}$$

[28] Note that, if there are no fixed site costs (i.e., if $C_s = 0$), then qn is a constant (=C/c) and the within-site variance component of equation 18 is a constant. In that case, increases in q would reduce the variance of the impact estimate without limit, and the optimal number of sites would be the maximum attainable, C/c, with a single sample member in each site.

ponent of the standard error of the impact estimate (see equation 18). The optimal value of q is that which just balances the gains from additional sites against the loss in sample size that results from additional fixed site costs.

To solve this problem, we must specify the values of the parameters in equations 18 and 19. In most cases, the overall budget C will be known and the cost parameters C_S and c can be estimated with reasonable accuracy. Moreover, the within-site variance of the experimental estimator can usually be estimated from existing data on the outcome of interest.[29] But the between-site variance of the experimental impact is generally unknown.

In some cases, however, it is possible to derive at least a proxy for the between-site variance of the impact from nonexperimental data. In the Health Insurance Experiment, for example, the experimenters used the variance of the difference in medical expenditures between insured individuals and uninsured individuals across regions of the country as a proxy for the between-site variance of the impact of cost-sharing on medical expenditures. More generally, there may be prior multi-site nonexperimental impact studies of the same, or a similar, intervention, from which an estimate of the cross-site variance (*i.e.*, the variance of site-specific impacts) can be derived.

If all else fails, the cross-site variance can be "guesstimated" as follows. Posit the range that you would expect to include 95 percent of all site-specific impacts. (In most cases, the lower bound of this range will be zero.) If site-specific impacts are normally distributed, this range will correspond to 1.96 standard deviations of the distribution of site-specific impacts. An estimate of the between-site variance of impacts, then, is approximately the square of one-half this range.

Whatever the source, if estimates of the parameters involved in equations 18 and 19 can be obtained, an estimate of the optimal number of sites can be obtained. Archibald and Newhouse (1988) provides an analytic solution for the minimization of equation 18 subject to the constraint in equation 19, developed by Carl Morris.[30] A more direct method is simply to compute the value of SE_I from equation 18 for every integer value of q between 1 and the maximum number of sites that could be supported by the experimental budget ($= C/C_S$) and choosing the value that minimizes SE_I directly.[31] This approach has the advantage of also showing the loss in power associated with a nonoptimal number of sites.

As in all design decisions, the quality of the result obtained from this procedure—*i.e.*, the likelihood that the choice truly maximizes the power of the design—depends on the quality of the information that went into the decision. As the forgoing discussion makes clear, in many cases only relatively low quality information about some of the critical parameters may be available. But some information is almost certainly preferable to none, since some decision must be made. Even fairly unreliable information about the critical parameters may be sufficient to establish the right order of magnitude.

Random Assignment of Groups of Individuals

Up to this point, we have assumed that the focus of policy interest is on program effects on individuals and that experimental subjects are randomly assigned independently. This will be the case in most social experiments. There are two situations, however, in which it may be necessary to randomly assign *groups* of individuals.

The first occurs when policy interest focuses on impacts at the aggregate level and random assignment of individuals is inconsistent with unbiased estimation of aggregate effects. Consider, for example, an experiment designed to test the effects of a worker training program on productivity at the plant level. In such an experiment, one cannot randomly assign individual workers to the training program; only if all the workers in the plant are subject to the same policy regime will we obtain unbiased effects at the plant level.[32] Instead, one would have to randomly assign *plants* to a treatment group, in which the program would be implemented plant-wide, or to a control group, in which it would not be implemented at all. All of the principles of sample design discussed in this paper would then apply if one simply defines the plant, rather than the individual worker, to be the unit of analysis.

---

[29] Under the null hypothesis, the within-site variance of the experimental impact estimate is just:

$$V_I = V_Y/n_t + V_Y/n_c,$$

where V_Y is the variance of the outcome and n_t and n_c are the numbers of treatment and control group members, respectively, in a single site. Existing data are usually available to estimate V_Y.

[30] See Morris (1974). In the case of the Health Insurance Experiment, the optimal number of sites was in the range of 4-9, for reasonable values of the parameters (Newhouse, 1993).

[31] For each value of q, equation 19 must be solved for n, in order to calculate SE_I.

[32] This is not to say that all workers in the plant must participate in the training in order for the estimates to be valid, only that none who would participate in an ongoing training program should be *artificially* excluded through random assignment.

It may, of course, be difficult to obtain the cooperation of a sufficient number of plants to generate the sample size needed for reliable estimates. Whereas experimental samples of thousands of individuals are quite common, it may be difficult to recruit more than, say, 10 or 20 plants to participate in an experiment. Whether such sample sizes would be sufficient depends on the minimum detectable effects attainable with that number of observations, which in turn depends on the variance of the outcome of interest *across plants*. Fortunately, aggregate outcomes are frequently much less variable than outcomes at the individual level. In any case, minimum detectable effects can be computed for this case just as for samples of individuals, using the number of plants as the sample size and the variance of the outcome of interest at the plant level.

Though experiments of this type are much less common than those in which individuals are randomly assigned, several have been conducted in recent years. For example, the National Home Health Agency Prospective Payment Demonstration randomly assigned 142 home health agencies to alternative Medicare reimbursement formulae, in order to test their effects on the use of care.[33] Similarly, the San Diego Nursing Home Incentive Reimbursement Experiement randomly assigned 36 nursing homes to an incentive payment system designed to reward facilities for achieving various admission, treatment, and discharge objectives, or to a control group that received only the regular Medicare reimbursements.[34]

A second case in which random assignment of groups arises occurs when policy interest focuses on program effects on individuals, but it is infeasible to randomly assign individuals. This might be the case, for example, if one were interested in the effects of alternative teaching methods on students' achievement, but for institutional reasons it was not possible to randomly assign students to different classes. In that situation, one might be forced to randomly assign whole classes to different experimental groups. Similarly, when the experimental treatment applies to the family as a whole the entire family must be randomly assigned as a group. This was the case in the income maintenance experiments and the Health Insurance Experiment, where all members of the family were assigned to the same negative income tax or health insurance plan.

Calculation of minimum detectable effects at the individual level is more complex in this case, because the standard error of the impact estimate depends on the correlation among the outcomes within the "clusters" of individuals who were assigned together. Specifically, if SE_{I*} is the standard error of the impact estimate for random assignment of N sample members individually, the standard error of estimate for a sample of the same size randomly assigned in clusters of n individuals is:

$$\mathbf{20} \qquad SE_{IC} = SE_{I*} \cdot \sqrt{1 + d\,(n-1)}$$

where d is the **intraclass correlation** of the outcome.[35] The intraclass correlation is a measure of the homogeneity of sample members, in terms of the outcome Y, within the clusters randomly assigned. Its values range from $-1/(n-1)$ to $+1$, with positive values indicating similarity among individuals within clusters and negative values denoting dissimilarity within clusters, *relative to the makeup of the overall population from which they were drawn*. When $d=0$, the clusters are just as heterogeneous with respect to Y as the overall population.

As can be seen from equation 20, when $d=0$ the "**design effect**" of cluster sampling (the term under the square root sign) is one, and the standard error of the estimate based on random assignment of clusters is the same as the standard error of estimate based on random assignment of single individuals. When d is positive (*i.e.*, when individuals within clusters are more similar than the overall population), the standard error of estimate is higher under random assignment of clusters. When d is negative (*i.e.*, when individuals within clusters are more dissimilar than the overall population), the standard error of estimate is lower. Thus, minimum detectable effects will be larger under random assignment of clusters than under random assignment of individuals if d is positive, and it will be smaller if d is negative. If $d=0$, minimum detectable effects will be the same under the two approaches. Unfortunately, in practical applications, the intraclass correlation tends to be postive; *i.e.*, individuals within clusters tend to be more similar than individuals drawn randomly from the overall population. Thus, cluster effects generally tend to *increase* the standard error of estimate.

Suppose, for example, that we randomly assign classrooms of 25 students each to alternative teaching methods. Further suppose that students have been assigned to classrooms in part on the basis of ability, leading to a positive intraclass correlation of grade point average (the outcome of interest) of 0.20. According to equation 20, clustering will increase the standard error of estimate, and therefore minimum detectable effects, by a factor of 2.4. That is, when we randomly assign classrooms rather than individual students,

[33] See Goldberg (1997).

[34] See Jones and Meiners (1986).

[35] See Hansen et al. (1965), or any standard text on sampling statistics, for a formal definition of the intraclass correlation.

the experimental effect on grade point averages would have to be 2.4 times as large to be detectable at a given level of statistical significance. In order to achieve the same power as a design in which individual students were randomly assigned, the sample size would have to be increased by a factor of 5.76 ($=2.4^2$).

As this illustrative example suggests, random assignment of groups can result in a substantial loss of power, relative to random assignment of individuals. The size of the loss will depend on the intraclass correlation of the specific outcome of interest and the size of the groups randomly assigned. In cases where random assignment of groups is the only feasible approach, it is therefore critical to estimate the minimum detectable effects attainable with the proposed sample size and design, taking cluster effects into account, to ensure that the design will yield estimates of sufficient power to be worthwhile.

References

Archibald, Rae W., and Joseph P. Newhouse. 1988. "Social Experimentation: Some Why's and How's." In *Handbook of Systems Analysis: Craft Issues and Procedural Choices*, ed. Hugh J. Miser and Edward J. Quade, pp. 173-214. New York: North Holland.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," *Evaluation Review,* Vol. 19:5, pp. 547-556.

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources* 21 (Fall): 606-39.

Goldberg, Henry B. 1997. "Prospective Payment in Action: The National Home Health Agency Demonstration." *CARING Magazine* XVI, no. 2: 14-27. February.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. 1965. *Sample Survey Methods and Theory. Volume 1: Methods and Applications.* New York: John Wiley and Sons, Inc.

Hausman, Jerry A., and David A. Wise. 1985. *Social Experimentation*. Chicago: The University of Chicago Press.

Jones, Brenda J., and Mark R. Meiners. 1986. "Nursing Home Discharges: the Results of an Incentive Reimbursement Experiement," *Long-Term Care Studies Program Research Report.* DHHS Publication No. (PHS) 86-3399. Rockville, Md.: U.S. Department of Health and Human Services, Public Health Service, National Center for Health Services Research and Health Care Technology Assessment. August.

Kish, Leslie. 1965. *Survey Sampling.* New York: John Wiley and Sons, Inc.

Morris, Carl. 1974. "An Estimate of the Optimal Number of Sites." Health Insurance Study memorandum SM-1093 (unpublished). Santa Monica, CA: Rand Corporation.

Orr, Larry L., Terry Johnson, Mark Montgomery, and Marie Hojnacki. 1989. *Design of the Washington Self-Employment and Enterprise Development (SEED) Demonstration.* Bethesda, MD: Abt Associates Inc. and the Battelle Memorial Institute.

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