Social Experiments in Education

Nuts and Bolts of Designing Study Samples: Chapters 5 and 6 (Review Versions)
Comments Welcome: Rmaynard@gse.upenn.edu

Rebecca Maynard and Larry Orr (with Nianbo Dong)
11/23/2009

This document contains drafts of two chapters from a book we are writing on Social Experiments in Education. We have benefitted in this work from the work of many others who have written on this topic, including Howard Bloom, Steve Raudenbush, Peter Schochet, and Steve Bell. We also have gained much from working with current and former doctoral students, especially Dionissi Aliprantis, Erica Johnson, and Nirav Mehta. Partial funding for this work was provided by the Institute of Education Science, for which we are grateful. We alone are responsible for the content.
Chapter 5
SAMPLE DESIGN

The...Stochastic Art...is the art of measuring as accurately as possible the probabilities of things so that in our judgments and actions we can always choose or follow that which seems to be better, more satisfactory, safer and more considered.

-- Jakob Bernouli, Swiss mathematician and discoverer of the law of large numbers, 1713

In this chapter, 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 five issues:

- Site selection and the external validity of the experimental impact estimates
- Sample size, statistical significance, and the power of the design
- Random assignment of groups of individuals
- 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.

SITE SELECTION AND THE EXTERNAL VALIDITY OF THE ESTIMATES

As noted in Chapter 1, 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 study sample must provide

---

1Bernouli, Jakob, *Ars Conjectand*, Translation by Sheynin (2005) downloaded 12_30_8 at [http://www.sheynin.de/download/bernoulli.pdf](http://www.sheynin.de/download/bernoulli.pdf). The law of large numbers states that by taking a sufficiently large sample, one can increase to any desired level one’s confidence that a sample statistic is within a stated amount of the corresponding population value. The law of large numbers underlies the entire modern approach to sampling. The term “stochastic” refers to a random process that generates different values, each with some probability; the individual values cannot be predicted, but the distribution of all observations usually follows a predictable pattern that can be described by statistical means.
unbiased estimates of the impact on the population of interest for policy.\textsuperscript{2} External validity is sometimes characterized as \textit{generalizability}—that is, externally valid estimates can be generalized to the population of interest for policy. In order to provide the most reliable guidance for policymakers, the results of social experiments should be both internally and externally valid. Suppose, for example, that we test a new reading curriculum on a sample that includes an atypically large proportion of students who are from non-English-speaking families. Even if the study yields perfectly unbiased impact estimates for this sample, the impact estimates may be very misleading indicators of the effect the new curriculum would have on a larger, more typical student population. 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.\textsuperscript{3} 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 produces a sample that does not differ systematically from that population. Thus, if the study sample is randomly drawn from the population of interest, internally valid impact estimates will also be externally valid—i.e., they will provide 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 high school students in the U.S., or even in a single state. Such a

\textsuperscript{2}As this definition implies, in order for impact estimates to be externally valid, the estimates also must be internally valid.

\textsuperscript{3}Random sampling should not be confused with random assignment to study condition (e.g., treatment or control group). In random sampling, a group of individuals is selected randomly from a larger population in order to obtain a sample for analysis that is representative of the population from which the sample was drawn. In random assignment, the study sample is divided randomly into two or more groups, each of which is subjected to a different program, policy or practice relevant to particular research questions.
sample would be spread so thinly over a large number of schools and geographic areas that the costs of administering the program, policy or practice under study across that number of location and of collecting data for the study sample would be prohibitive. Instead, study samples are generally clustered in a small number of “sites”—schools, communities, or school districts.

It is still possible to obtain a random sample of the overall population (e.g., all high school students in the U.S. or in a given state) if the study sites are randomly selected from all sites in that population and study participants are then randomly selected from the population of interest within each site. Such a sample design is known as a multi-stage random sample. This type of sampling procedure was used in the National Head Start Impact Study (NHSIS) (Puma et al. 2005). In that study, a sample of 84 local Head Start grantees and delegate agencies was randomly selected from a sampling frame of approximately 1,530 program grantees and delegate agencies that combined contained 85 percent of the national population of program participants. As the result of an intensive site recruitment effort, all 84 sites selected at random for the study agreed to participate. Subsequently, a sample of 4,667 Head Start applicants was selected randomly from among all eligible applicants in the 84 sites to participate in the study sample; the applicants in the study sample were then randomly assigned to the program group, which was eligible for Head Start beginning immediately, or to a control group, which was excluded from Head Start participation for one year.

One of the potential barriers to obtaining a representative sample of the population of interest is the need to obtain the cooperation of local school administrators or program operators. School administrators and program operators typically resist participating in social experiments

---

4 Grantees whose programs fully saturated their communities (i.e., served all interested families/children, with little potential to expand the applicant pool enough for the evaluation control group without substantially reducing total enrollment) were excluded from the sampling frame. Thus, the sample actually represents the national population of Head Start programs facing excess demand.
for a variety of reasons, including the added burden of experimental sample intake and random assignment procedures, fear of disruption of ongoing school or program activities, and ethical concerns about denial of service to those assigned to the business as usual control group.\footnote{In a subsequent chapter, we discuss ways to address these concerns.} If high proportions of sites selected for the study sample refuse to participate, selection bias can creep into the impact estimates. 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 for the intended target population.

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 been conducted on \textit{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 or school, sites have been selected by sponsoring agencies on political grounds. Sometimes, such selections are a \textit{fait accompli} before the research team has been selected.

At best, convenience samples of sites leave the experimenter with limited 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 study sample within a self-selected set of sites, they inject the very selection bias that social experiments are intended to avoid, albeit in a more subtle manner. In most cases it is, of course, possible to compare the characteristics of the study sites and of the analysis sample with those of
the broader populations from which they were drawn. Such comparisons can identify ways in which the study 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.

**Purposive selection** of sites that are well-matched to the population of interest in observable characteristics is a commonly used alternative to both the random selection of sites and the use of convenience samples. For example, the recently completed study of Teach for America drew its study sample purposively from a large number of dispersed placement communities throughout the United States (Decker et al., 2004). Another recent study of professional development in early reading instruction for elementary school teachers selected its study sample through a careful, purposeful screening of schools that met particular characteristics (Garret et al. 2008).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 (e.g., ten or fewer) because in small samples sampling error can create large differences between the sample and the population from which it was drawn. Purposive selection directly controls such differences in some 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

---

6 An even more elaborate example of purposive sampling was employed in 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 (see Orr et al. 1989).
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 evaluator still has to convince local school administrators 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, economic, or academic 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 selecting a set of schools in which to test a new curriculum, one might try to pick some schools with high average test scores levels and some with low average test scores, some schools in areas with high poverty rates and some in areas with low poverty rates, and so forth. Such an approach can help researchers to understand how the impact of the program, policy or practice under study varies with these conditions. But if these conditions do influence the 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 impact of the intervention based on the study sample will not be an unbiased estimate of the average impact expected for the broader population. To obtain an unbiased estimate of what the impact would be for the broader population, one would have to “reweight” the sample to reflect the composition of that population in these dimensions, prior to generating the impact estimates. 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 that were 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 Chapter 3,
this means that the sample selection and random assignment process within the experimental
sites must be designed to yield a sample that is reflective of the relevant population—whether
that is the overall target population, eligible applicants, or potential participants. It also means
that the implementation of the experiment should disturb the existing educational process as little
as possible.

In practice, it is often extremely difficult to achieve an externally valid study sample.
Evaluators often lack the resources needed to recruit a truly representative sample of sites and to
induce local school administrators or program operators in all of the selected sites to participate
in the experimental-design study. The results of efforts to achieve a study sample that is
representative of the relevant population group must be judged both in comparison to the ideal of
a perfectly representative sample and 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 school administrators, experimental evidence from
even a 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 for estimating impacts (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.

Researchers sometimes distinguish between “efficacy trials” and “effectiveness trials.”
In efficacy trials, the objective is to determine whether the intervention has the intended effect on a small scale, under optimum conditions, whereas the aim of “effectiveness trials” is to estimate the impacts of the intervention in a more representative (i.e., externally valid) setting and with a representative study sample. Efficacy trials generally are conducted with convenience samples of sites or, at best, purposively selected sites; consequently, it is not possible to generalize the results to the larger population. The results of such trials are “fixed effects” estimates that apply only to the sample on which they were estimated. Effectiveness trials are conducted in a random sample of sites or settings; their results, which are referred to as “random effects” estimates, can be generalized to the population from which the sites were drawn. (These distinctions are discussed in more detail in a later chapter.)

Ideally, efficacy trials are conducted first, and those interventions that are found to be efficacious are then tested in effectiveness trials. This strategy avoids expending resources on the more expensive effectiveness trials of interventions that do not produce the intended effects even under small scale, highly favorable conditions. At the same time, it still requires that interventions be proven effective for a representative population, in real world educational settings. Unfortunately, all too often positive findings in efficacy trials are taken as evidence of effectiveness and are not confirmed with trials in more representative settings.

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 experimental-design study is being implemented. Indeed, if the policy tested is one that may be adopted by local school systems or community service providers, there may be many different contexts in which the policy could be considered for adoption. It is, therefore, critical that the nature of the intervention being studied and the sample selection procedures be carefully documented, so that future policymakers will
know how closely their situation corresponds to the program, policy, or practice was tested and
the target population it was tested on. As we will discuss in a later chapter, it also will be helpful
if, in the analysis, the evaluator measures the extent to which impacts vary across sites and/or
whether they vary with site conditions. The less the impacts vary across sites, or the more
predictable that variation is, the more applicable the study findings will be to localities whose
characteristics differ from those characterizing the average site and sample member in the study
sample. Of course, the ability to measure variation in impacts across sites is limited by the
amount of diversity among the study sites. Therefore, it is highly desirable, even in efficacy
trials, to select sites that represent diverse student populations and institutional contexts, even if
those sites are not fully representative of the larger population of interest.

SAMPLE SIZE, STATISTICAL SIGNIFICANCE, AND THE POWER OF THE DESIGN

As explained in Chapter 2, 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, tell us whether we can be confident that the impact is greater
than zero, and specify a confidence interval around the estimate that we can be reasonably
certain includes the true impact. In designing an experiment, one of our central objectives is to
ensure that the confidence with which we can say whether the program had an impact of a given
size is great enough, and the interval within which we can bracket the true impact is narrow
enough, for policy purposes. For any given true impact of the intervention, “the power of the
design” is the probability that the experiment will detect that impact as statistically significant—
i.e., that it will reject the null hypothesis of no impact. In this section, we discuss the factors that
determine the power of the design and the criteria for deciding what power one should design the experiment to achieve.

Setting the Objectives of the Design

In designing an experiment, the researcher must begin by answering several fundamental questions that define the objectives of the design. First, how large an impact do we want to be able to detect? The answer to this question should be based on the size impact that would be meaningful for policy or practice—i.e., how large would the impacts of the intervention have to be for it to be considered for adoption as a regular policy or program? Designing the experiment to detect smaller effects would entail an unnecessarily large sample and needless expense, since such effects would not affect policy or practice. But, if the experiment is incapable of detecting effects that are meaningful for policy or practice, the researcher risks missing impacts that would represent a meaningful improvement over the current educational program, policy or practice. Such experiments are commonly described as “under-powered”.

We call the smallest effect that is meaningful for policy the “minimum relevant-size impact,” or the MRI. At the outset of every experimental-design study, the researcher should specify the MRI for each central outcome to be measured and, following the procedures outlined in this chapter and the next, design the study to provide a high probability that, if the intervention being studied really has an impact equal in size or larger than the MRI, the study will detect it. Because what is meaningful for policy or practice must ultimately be decided by educational policy makers or practitioners, the MRI should be specified in consultation with representatives of the policy and/or practice community.

The second question the researcher must answer is, how confident do I need to be that I can detect a true impact equal to the MRI—i.e., if the true impact is at least as large as the MRI,
how high do we want to set the probability that the study will reject the null hypothesis of zero effect? The higher we set that probability—the power of the experiment—the more confident we can be that the experiment will not fail to detect a meaningful impact of the intervention when, in fact, there really is one. But setting the power of the experiment high is not costless; as we will see, higher power demands larger samples and, therefore, greater costs of conducting the study.

The third question the researcher must answer is, how great a chance am I willing to take that the study will conclude that the intervention had a real (nonzero) impact when, in fact, it did not—i.e., that it will yield a “false positive” result? The risk of a false positive result is given by the significance level of the test of the null hypothesis. Again, reducing the chance of a false positive is not costless—other things equal, a smaller risk requires a larger sample.

The conventional answers to these last two questions have been to strive for 80 percent power to detect true impacts deemed to be equal or greater than the MRI and 5 percent significance levels for the tests of the null hypothesis. But the researcher should not just blindly accept these values; s/he should consider these risks within the context of their specific study. And, as we will see below, for any given sample size (i.e., for any given study budget), there is a trade-off between power and the risk of false positive findings. We can reduce the risk of missing a real nonzero effect if there is one by accepting a higher risk that we will conclude that there is a nonzero effect when there is not one.

Measuring the Power of the Design

As noted above, the power of the study 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 new curriculum on student test scores, and we have specified an MRI of 5 points. If the true impact of the curriculum is positive (i.e., if it does in fact increase
test scores), we would like the test of statistical significance of the impact estimate from the study 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.

Exhibit 5.1 shows how the power of the design can be calculated for one specific value of the true impact, say a 5-point increase in test scores. That is, it shows how to calculate the probability that, if the true effect of the experimental program is to increase test scores by 5 points—exactly the amount of the MRI—the impact estimate from the study will be statistically significantly greater than zero (i.e., that the null hypothesis will be rejected). 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. (Note that one must specify the significance level of the test in order to calculate the power of the design.) As explained in Chapter 2, if the experimental estimate falls within the critical region, the null hypothesis of zero impact is rejected. 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 5 points.

[Exhibit 5.1]

The answer to that question is given by the sampling distribution of the experimental estimate when the true impact on student test scores is 5 points; this is the distribution to the right in Exhibit 5.1. This distribution is centered on 5 and its shape is determined by the standard error of the estimated difference in means between the intervention and control groups. The probability that the experimental estimate will fall in the critical region if the true impact is 5 is given by the area under this curve to the right of the critical value $I_c$. This probability is the
power of the design for a true impact of 5 points—in this illustrative example, 70 percent. That is, with the design represented here, there is a 70 percent chance that we would reject the null hypothesis of zero effect when the true effect is 5 points. (Later in this chapter, we will explain how the numerical value of this probability is calculated; here, our interest is in its conceptual derivation.)

As this example makes clear, the power of the design depends on the shape of the two sampling distributions in Exhibit 5.1. And as noted in Chapter 2, the shape of the sampling distribution of the experimental estimate depends on the size of the study sample. Other things equal, the larger the study sample, the more tightly the sampling distribution will be clustered around its mean.

Exhibit 5.2 shows what happens to the power of the study design when the sample size is increased from that depicted in Exhibit 5.1. For two reasons, increasing the size of the experimental sample increases the power of the design. First, the tighter sampling distribution around the true impact of 5 points increases the probability of the experimental estimate exceeding any value to the left of 5, including the critical value for the test of significance. Second, the tighter sampling distribution around the null hypothesis of zero impact lowers the critical value for the significance test, which also increases the proportion of the area under the sampling distribution around 5 points that lies above the critical value (compare the shaded area under the right-hand curve in Exhibit 5.2 with the corresponding shaded area in Exhibit 5.1).

[Exhibit 5.2]

Power and the Significance Level of the Test

A second way in which we could increase the power of the design would be to change 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, $I_c$, and, thereby, increase 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 5 points.

However, this change in 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: (1) 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 (2) 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 established by the significance level of the test of the null hypothesis. The probability of the latter error is one minus the power of the study design to detect true impacts that are at least equal in size to the MRI.

In the design depicted in Exhibit 5.1, there is a 5 percent risk of falsely concluding that the program effect was positive when, in fact, it was zero, and there is a 30 percent risk of failing to detect a true 5-point impact on test scores. In making the trade-off between these two risks, researchers typically accept a higher risk of failing to detect a nonzero impact than of falsely concluding that the impact 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. In contrast, if we fail to detect true positive

---

7 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, or a “false positive”. 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.
impacts, we miss an opportunity to implement an effective program, but we do not waste large sums of money.8

However one views these risks, it is important to make the trade-off explicit. 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 or lowering the significance level to achieve a better balance between the two types of risk. If it is not possible to obtain a sufficiently large sample to yield adequate power, it may be appropriate to forgo conducting the study altogether.

Minimum Detectable Effects and the Optimal Sample Size

In the previous example, we took sample size and significance level as given and solved for the power of the design to detect a true impact equal to the MRI. In practice, education researchers typically specify the desired power to detect effects at a specified significance level of the test. Given these two parameters and information about the outcome variable and the sample size and its allocation between the intervention and control groups, it is possible to determine what the “minimum detectable impact” (MDI) will be. The MDI is the smallest true impact that a given design can detect as statistically significant at the specified levels of power and significance. As we will see, the MDI is a critical to determining the optimal sample size for the experiment.

We begin by showing how the MDI can be computed. We then show how it can be used

---

8Similar reasoning would apply to a study of an ongoing program, policy or practice. A false positive finding would result in the continuation of current practice and a waste of resources, whereas a false negative would result in terminating the practice that is beneficial. The social cost of the latter will depend on the alternative use of the resources freed up when the beneficial practice is terminated.
to set the optimal sample size.

Suppose that in the case of the curriculum test discussed above, we are willing to take a 10 percent risk of rejecting the null hypothesis of zero when it is in fact true, but we want to be 80 percent certain that if there are true impacts meeting an established threshold, the study will conclude that the intervention was effective. If we know the variance of the outcome test score measure, it is quite straightforward to calculate the minimum detectable impact (at the established standards of statistical power and significance levels) for any particular total sample size and allocation between the intervention and control groups. Exhibit 5.3 illustrates the critical pieces of information needed to compute minimum detectable impacts. 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 (two-tail test).

[Exhibit 5.3: Calculation of Minimum Detectable Effects]

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 depicts the sampling distribution of the estimated impact of the intervention. In order to have an 80 percent probability of rejecting the null hypothesis of zero effect when the true impact is 5 points, 80 percent of the area under this sampling distribution must lie to the right of the critical value \(I_C\), which in this example = 1.64*\(SE_i\). This will be the case when the mean of the distribution of impact estimates lies .84 standard errors above the critical value. Consequently, there is an 80 percent probability that, if the true impact is equal to 2.48 times the standard error of the impact

---

9 We show the derivation for the minimum detectable positive impact. With a two-tail test, the minimum detectable negative impact can be derived in the same way; since the sampling distributions are perfectly symmetric around zero, the negative MDI is just the negative of the positive MDI.

10 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.
estimate, the study findings will be at least equal to $I_C$ (1.64 times the standard error of the estimate) and, thus, be judged to be statistically significantly different from zero. This is the smallest impact that can be detected with 80 percent power at a .10 significance level. Thus, the MDI for these standards of power and significance is:

\[
\text{MDI} = 2.48 \times \text{SE}_I
\]

(5.1)

\[
= 2.48 \sqrt{V_Y / n_t + V_Y / n_c}
\]

(5.2)

where:

\[
V_Y = \text{the variance of the outcome measure}
\]

(5.3)

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 (two-tail test), the constant would be 2.80, rather than 2.48. Thus, more generally:

\[
MDI = M_C \sqrt{V_Y / n_t + V_Y / n_c}
\]

(5.4)

where $M_C$ is a constant that reflects the multiplier applicable to the chosen levels of power and significance.

As is clear from equation 5.4, all that is required to compute the MDI for any given value of $M_C$ 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 student test scores can be computed on the basis of data routinely collected by schools in their regular testing
programs. Data for other outcomes, such as attendance or high school completion rates, can usually be obtained from existing administrative records or data collected as part of other evaluations. 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.

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

An important property of experimental designs is readily derived from Equation 5.4:

For any given division of the sample between treatment and control groups, minimum detectable impacts are inversely proportional to the square root of the overall sample size. To see this, let \( P_t \) be the share of the total sample \( N \) allocated to the treatment group and \( P_c \) be the share of the total sample allocated to the control group. Then Equation 5.4 can be written as:

\[ \text{Equation 5.4} \]

---

11 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, such as high school completion, tested proficient or better, or entered college), the variance of the outcome is \( p(1-p) \), where \( p \) is the sample mean or the proportion of the sample taking on the value 1. The expression \( p(1-p) \) is maximized when \( p = .5 \). Therefore, even if the mean is unknown, one can compute the “worst-case” minimum detectable impact estimate for dichotomous outcomes using the assumption that \( p = .5 \).

12 Later in this chapter we discuss the optimal allocation of the sample between the treatment and control groups. For now, we simply note that, from the standpoint of statistical power, the optimal sample allocation in a simple two-group design like this one is equal numbers of observations in each group. As we will see in Chapter 6, however, operational considerations may dictate departures from this allocation.
Thus, for example, doubling the sample reduces the MDI by a factor of $1/\sqrt{2} = .71$; in order to halve the MDI, one would have to quadruple the sample.

As is also clear from Equations 5.4 and 5.5, if we specify the desired MDI and the allocation of the experimental sample between the intervention and the control group, we can solve for the sample size that will yield any particular MDI at the significance and level of power embodied in the multiplier ($M_c$). As discussed at the outset of this section, the smallest impact we would like to be able to detect with high confidence is the minimum relevant-size impact, or MRI. The optimal sample size is therefore the sample size that satisfies the condition:

$$MDI = MRI$$

(5.6)

If we set the MDI equal to the MRI and specify the sample allocation between the two experimental groups, equation 5.5 can be solved for the optimal overall sample size:

$$N = (M_c^2 / MDI^2)(\sqrt{V_y / P_T} + \sqrt{V_y / P_C})$$

(5.7)

Thus, for any given sample allocation between the intervention and the control group (a topic to which we will return), the answers to the three questions posed at the beginning of this section—the levels of the MRI, significance level, and desired power—determine the sample size needed for the experiment.

Minimum Detectable Impacts in Effect-size Units (MDI-ES) and Power of the Design

As just discussed, specification of the MRI, along with the desired power and significance level for the experiment, determine the optimal sample size and the ability of the experiment to detect the true effects of the intervention. Setting the MRI is, however, a judgmental exercise that
requires in-depth knowledge of education practice and policy. The problem of setting the MRI is especially difficult when the outcome of interest is measured in units that have no intrinsic intuitive meaning. For example, numerical values of the social or emotional scales sometimes used to measure young children’s development often have no obvious interpretation. In such cases, it is helpful to convert these scores to some measure of the variation of the scale across the population of interest.

One metric that can be used for this purpose is the standardized mean difference, or “effect size.” Effect size is the estimated impact (i.e., the difference between the intervention and control group means) expressed in standard deviations of the outcome variable, rather than natural units. An effect size, then, tells us how far the intervention moves the treatment group mean relative to the distribution of outcomes for the control group.\(^{13}\) The “minimum detectable effect size” (MDI-ES) is the MDI expressed in standard deviations of the outcome variable.\(^{14}\)

Suppose, for example, that in the example described above the standard deviation of test scores was 15; in this case, an MDI of 5 test score units would translate into an MDI-ES of .33, or one third of a standard deviation. The MDI-ES is calculated the same way the MDI is calculated, except that the relevant points in the sampling distribution (see Exhibit 5.3) are measured in standard deviation units, rather than natural units. This means that the MDI-ES can be calculated without knowledge of the actual variance of the outcome variable; that value is only needed to relate the MDI-ES back to natural units. Of course, if one doesn’t know the variance of the outcome, it is impossible to interpret the MDI-ES, and therefore to make sensible

---

\(^{13}\) In some cases, effect sizes are based on the standard deviation of the outcome for the overall study sample, including the intervention group, the national population, or some other reference population, rather than its standard deviation in the control sample alone.

\(^{14}\) In the literature, the term “minimum detectable impact” (MDI) or “minimum detectable effect” (MDE) is often used both when the effect size is measured in natural units and when it is measured in standard deviation units. For clarity, we use the term MDI-ES for the latter here.
design decisions. We return to the interpretation of the MDI-ES below.

Effect sizes are also sometimes used to compare impacts or power across different experiments. For example, a funding agency may set a single MRI in effect size units to standardize the power of the experiments it funds, or may use the MDI-ES measures to compare power across them. The MDI-ES has also been used to compare the power of studies whose central outcomes are measured using different tests or even constructs (e.g., credit accumulation versus test scores).

Measurement of impacts in effect size units has become standard practice in education research and evaluation. There are even widely adopted conventions for the ranges of effect sizes that constitute “small”, “moderate”, and “large” impacts. However, effect size measures have several distinct weaknesses that make them poor choices as indicators of the size of impacts or the power of study designs. First, while they may facilitate the interpretation of impact estimates for abstract constructs such as measures of social and psychological well-being, in many other contexts they are not as intuitive as measures expressed in natural units, such as percentile rank or months of educational progress, especially for non-researchers. Certainly, most educators or policymakers will not know what to make of an effect size of .33 without further explanation. They would be much more comfortable assessing the magnitude of an impact that represented, say, an improvement of 4 months in reading ability.

Second, a given effect size may correspond to very different size impacts in natural units for different outcomes or student populations. This point is illustrated in Exhibit 5.4, which compares the impacts in several different natural metrics that correspond to an effect size of .2.

---

15 Another application in which impacts are almost universally measured as effect sizes is meta analysis, in which estimated impacts – sometimes on different outcomes – are, in effect, averaged across a number of experiments. The caution here is that effect size measures are very sensitive to the choice of a reference population for the standard deviation estimate used in translating outcomes measured in natural units to effect size measures.

16 See Cohen (1988) for the original source of these descriptors.
generally regarded a small effect size.\textsuperscript{17} Several important patterns can be seen in the exhibit. First, the equivalence between effect size and natural units is highly dependent on grade level for some metrics, but not for others. For example, an effect size of .2 is 56 percent of the mean annual reading gain between third and fourth grades, but 333 percent of the mean gain between eleventh and twelfth grades. In general, a given effect size represents a greater percentage of average annual performance gains at higher grade levels (i.e., average annual reading gains decline with age, relative to their standard deviation).\textsuperscript{18} In contrast, an effect size of .2 is a much more stable percentage of the white-black reading gap in reading – 24 percent at grade 4 vs. 30 percent at grade twelve.

[Exhibit 5.4]

A second pattern that is evident in Exhibit 5.4 is that, within a given grade level, the same effect size may seem larger or smaller when expressed in different natural units. For example, a reading gain of over three-quarters of a year between 7\textsuperscript{th} and 8\textsuperscript{th} grades may seem much more substantial than elimination of a quarter of the white-black reading gap. Similarly, for eighth graders, an effect size of .2 represents 111 percent of the performance gap between average and “weak” schools (i.e., those at the 50\textsuperscript{th} and 10\textsuperscript{th} percentiles of the distribution of schools within school districts), but only 30 percent of the performance gap between students who are eligible for free or reduced price lunch and those who are ineligible.

There are also serious issues with respect to the choice of the standard deviation used to standardize the effect size. Some studies use the standard deviation of the national population on which the outcome measure was normed; when this distribution is not available, and sometimes when it is, researchers use the standard deviation of the study sample distribution. Analysts may

\textsuperscript{17}The values in Exhibit 5.4 and many of the ideas discussed in this section, are taken from Bloom et al. (2008). See the footnotes to the exhibit for specific sources within that paper.

\textsuperscript{18}The same is true of nationally-normed tests in math, science, and social studies. See Bloom et al. (2008).
also choose to standardize on the standard deviation at the student level or the school level. There is no clear consensus in the literature as to which of these is the most appropriate measure, but clearly the choice of standard deviation used in computing effect sizes matters greatly. Sample standard deviations tend to be smaller than those for national populations, because samples tend to be more homogeneous than the corresponding national population. Thus, for any given impact estimate, effects sizes will be larger if the sample standard deviation is used than if the national standard deviation is used. Student-level standard deviations may be several times as large as those measured at the classroom or school level, where much of the variation in outcomes across students has been averaged out. For any given impact, this will lead to smaller effect sizes at the individual student level than at more aggregated levels.

Even within the majority of studies that use the sample standard deviation at the student level to standardize effect sizes, there are important definitional differences. Some use the standard deviation for the control group, while others pool the treatment and control groups for this purpose. Some use within-district or within-state standard deviations, while others use measures that include the variation across districts or states. Perhaps the most important difference is between those that use regression-adjusted standard deviations vs. those that use unadjusted measures. All of these differences will give rise to differences in the effect size calculated from a given experimental impact.

Finally, it is clear from Exhibit 5.4 that the conventional characterization of an effect size of .2 as “small” is not sensible – e.g., elimination of the entire reading gap between an average school and one at the 10th percentile of the distribution within a school district would not be considered a small impact by most educators. Much the same is true of several other outcomes in the Exhibit.
These issues suggest to us that effect sizes are not a good metric for the precision of an experimental design. Rather than using effect sizes, we recommend that, wherever possible, MRIs and MDIs should be measured in natural units appropriate to the grade level and objectives of the intervention tested. Thus, for example, if the intervention is a new curriculum intended to improve the mathematical ability of low-performing sixth-grade students, we would express the MRI in terms of the fraction of the performance gap between high- and low-performing students or average annual progress in math for low-performing students in grade 6. Expressed in these terms, it will be much easier for the funding agency and/or policymakers to decide how small an effect would be meaningful for educational policy – i.e., to set the MRI. This approach also forces evaluation funders and designers to be clear about the objectives of the intervention at the outset.

**RANDOM ASSIGNMENT OF GROUPS OF INDIVIDUALS**

Up to this point, we have assumed that the focus of policy interest is on program effects on individual students and that experimental subjects are randomly assigned one at a time. In many evaluations of educational interventions, however, it may be necessary to randomly assign *groups* of individuals.

One such case occurs when policy interest focuses on impacts at the school or district level and random assignment of individuals is inconsistent with unbiased estimation of effects at those levels. Consider, for example, an experiment designed to test the effects of a school-wide reform intended to improve achievement across the board, such as Success for All (Madden et al. 2003). Because such reforms must be implemented at the school level, and all students in the school will be affected by them, the unit of random assignment for studies of such interventions
must be the school. Even if it were possible to randomly assign students to schools with and
without the intervention, individual random assignment would not match the schools that
implemented the intervention with the control schools in other respects, such as quality of
teachers, physical facilities, and administration. To obtain a control group of schools that are
well-matched to the intervention schools, the unit of random assignment must be the school.

As we discuss in Chapter 6, the analysis of such an experiment could be done at either the
school level or the individual student level. If schools are the unit of analysis, the impact
estimate will be simply the regression-adjusted difference in mean outcomes between the
intervention and control schools. In that case, all of the design and analysis principles discussed
above and in previous chapters apply, with the school rather than the individual student as the
unit of analysis.

It may, of course, be difficult to obtain the cooperation of a sufficient number of schools
to generate the sample size needed for reliable estimates based on school-level means. Whereas
experimental samples of thousands of students are quite common, it may be difficult to recruit
large numbers of schools to participate in an experiment. The number of schools required would
depend, of course, on the MRI, as well as the variance of the primary outcome of interest across
schools. Fortunately, aggregate outcomes are usually much less variable than the same outcomes
at the individual level. In any case, MDIs can be computed for this case just as for samples of
individuals using Equation 5.5 above, with the relevant sample size units being the numbers of
schools assigned to the intervention and control condition and the relevant variance estimate
being that pertaining to school-level outcome measures.

A second case in which random assignment of groups arises occurs when policy interest
focuses on program effects on students, but it is infeasible to randomly assign individual
students. 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 this situation, conducting a rigorous, experimental-design study might require that we randomly assign whole classes to the intervention condition or to business as usual.

Calculation of MDIs at the student 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 students who were assigned together. Specifically, the standard error of estimate for a sample randomly assigned in \( N \) clusters of \( N_C \) students each is:

\[
SE_c = SE_{i}\sqrt{1 + ICC(N_C - 1)},
\]  

(5.8)

where \( SE_i \) is the standard error of the impact estimate that would be obtained if \( N \times N_C \) students were randomly assigned individually to experimental condition; \( N \) is the number of schools in the study sample; \( N_C \) is the number of students per school; and \( ICC \) is the *intraclass correlation* of the outcome.\(^{19}\) The ICC is a measure of the homogeneity of sample members, in terms of the outcome \( Y \), within the clusters randomly assigned. Its values range from \(-1(N_C-1)\) to \(+1(N_C-1)\), with positive values indicating similarity among individuals within clusters and negative values denoting dissimilarity within clusters, *relative to the makeup of the overall sample*.\(^{20}\) When \( ICC = 0 \), the clusters are just as heterogeneous with respect to \( Y \) as is the overall sample.

As can be seen from equation 20, when \( ICC = 0 \) the “cluster effect” (the term under the

\(^{19}\)See Hansen et al. (1965), or any standard text on sampling statistics, for a formal definition of the intraclass correlation.

\(^{20}\)Note that the lower bound of the intraclass correlation approaches zero as the cluster size \( N_C \) becomes large.
square root sign) is zero, and the standard error of the estimates based on random assignment of clusters is the same as the standard error of estimate based on random assignment of the same number of individual students. When the ICC is positive, the standard error of estimate is higher under random assignment of clusters and when the ICC is negative, the standard error of estimate is lower. Thus, MDIs will be larger under random assignment of clusters than under random assignment of individuals if the ICC is positive and smaller if d is negative. If the ICC = 0, the MDIs will be the same under the two approaches. Typically, we expect students within schools to be more homogeneous than the overall student population (i.e., we expect the ICC to be positive), so that cluster designs have larger standard errors and, therefore, lower power than designs with random assignment of the same number of students.

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, or simply that different schools serve neighborhoods with students of different ability, leading to a positive intraclass correlation of test scores of .10. According to Equation 5.8, the cluster effect will increase the standard error of estimated impacts on test scores, and therefore MDIs, by a factor of 1.84. That is, when classrooms are randomly assigned rather than individual students, the true effect on test scores would have to be 1.84 times as large as that under a design where the same number of students were individually assigned to intervention or control condition in order to be equally likely to detect it at specified levels of statistical power and significance. As a rule of thumb, for any set standards of power and significance, the total number of analysis units (NT = N*Nc) required increases with the square of the cluster effect. (Recall that the standard error of the impact estimate varies in proportion to the square root of sample size, as shown in Equation 5.5 above). Thus, in this example, to fully
offset the effects of clustering at the classroom level it would be necessary to triple the sample size.

As this illustrative example illustrates, random assignment of groups of students can result in a substantial loss of power, relative to random assignment of the same number of individual students. 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. While both of these factors will vary with the specific sample population and outcome of interest, it is possible to provide some fairly general guidelines as to what to expect in education data. Hedges and Hedberg (2007) provide empirical values of the intraclass correlation coefficient at the school level, for standardized tests of math and reading at each grade from kindergarten through 12th grade, based on national data. For both reading and math, and both for all schools and for low socioeconomic status schools, intraclass correlations were about .11, with somewhat higher values in grades K-8 and slightly lower values in high school. The authors also report that the average intraclass correlations in seven published studies that had cluster-randomized small numbers of schools were .17 for both math and reading.

The effect of correlations of this order of magnitude on precision will depend on cluster size. For clusters of 25 students, intraclass correlations of .10 to .17 will yield cluster effects of 1.84 to 2.25; for clusters of 100 students, the cluster effect will be 3.30 to 4.22. In order to obtain the same MDIs as a sample of individually assigned students, the latter effect would require a clustered sample containing roughly 18 times as large in terms of the units for analysis.

Of course, when outcomes are measured with administrative data, it may be quite feasible to obtain outcome data for very large numbers of students. But when surveys and/or special

---

21 Note that cluster size is defined as the number of students from a cluster that are included in the analysis sample. This need not be all the students in the unit randomly assigned. If, for example, schools are randomly assigned, the evaluation might randomly sample a subset of the students in each school, to reduce data collection costs.
testing are required, the costs of data collection will be high. In such cases, it is worth considering increasing the number of clusters and reducing the number of observations included in each cluster of the study sample. Chapter 7 provides guidance in the analysis of data from studies using cluster random assignment to ensure correct measurement of standard errors of estimate, taking cluster effects into account.

**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 students who are exposed to the intervention and not all of those randomly assigned participate in the program. As we shall see, the design of the random assignment process may affect the proportion of those randomly assigned who participate in the intervention, and this in turn will affect the power of the design to detect impacts on participants.

As shown in Chapter 2, under the assumption that the program had no effect on nonparticipants, an unbiased estimate of program impact on participants (IP) can be obtained by dividing the estimated impact on the overall treatment group (IT) by the participation rate among those randomly assigned (r):

\[ IP = \frac{IT}{r} \]  \hspace{1cm} (5.9)

The standard error of the estimated impact on participants (SE\textsubscript{IP}) can be derived from the standard error of the estimated impact on the total study sample (SE\textsubscript{T}) by the same procedure:\textsuperscript{22}

\[ SE_{IP} = \frac{SE_T}{r} \]  \hspace{1cm} (5.10)

\textsuperscript{22}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.
For any given sample design, the minimum detectable impact on participants (MDI_{IP}) can be expressed as:

\[
MDI_{IP} = M_C \cdot SE_{ip} = M_C \cdot SE_T / r, \quad (5.11)
\]

Thus, the minimum detectable impact on participants depends not only on sample size and the variance of the outcome measure, which determine the standard error of the impact estimate (SE_T), the multiplier implied by the established power and significance standards (M_C), and any cluster effect (all of which determine SE_T); it also depends on the participation rate among those randomly assigned (r). The lower the participation rate, the larger will be the minimum detectable effect on participants. Alternatively, for any given MDI_{IP}, the statistical power will decrease as the participation rate increases.

The principal way in which the evaluator can influence the participation rate and, thereby, improve the power of the design, is through the design of the random assignment process. Exhibit 5.5 shows the intake process for a voluntary summer school program targeted at low-achieving students.\(^{23}\) As can be seen in the exhibit, students assigned to the treatment group may fail to participate in summer school for various reasons and at several points in the process. Since the participation rate (r) in equations 5.9 - 5.11 is defined as the proportion of those randomly assigned to the treatment condition who participate in the program, this means that the later in the intake process random assignment is conducted, the higher will be the expected rate of participation in the summer school. Thus, the principal way the experimenter can increase the participation rate and reduce MDEs is by conducting random assignment as late in the intake process as possible.\(^{24}\)

\(^{23}\)This exhibit is adapted from Chapter 4, Exhibit 4.3.

\(^{24}\)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.
Consider, for example, the choice between randomly assigning all students who apply for the program regardless of eligibility status, only eligible students, or only eligible students who volunteer. Suppose that 20 percent of students do not meet the eligibility criteria and that only three-fourths of the eligible students will elect to participate in the program if offered the opportunity. This means that for every 100 applicants, there are only 80 eligible applicants and that, of those 80, 60 would ultimately participate in the program if all were offered the chance to do so. The participation rate among all applicants would be .60 (= 60/100), while that for eligible applicants would be .75 (= 60/80). This means that, for equal size samples, the MDI for a study sample of all applicants (MDI_{IP,A}) would be 25 percent larger than that for a comparable size study sample comprised of eligible applicants (MDI_{IP,E}). Specifically, the ratio of MDIs can be expressed as follows:

\[
\frac{\text{MDI}_{IP,A}}{\text{MDI}_{IP,E}} = \frac{(MC\cdot SE_T/r_A)/(MC\cdot SE_T/r_E)}{r_E/r_A} = r_E/r_A = .75/.60 = 1.25 \quad (5.12)
\]

Similarly, random assignment of all students, including those who will not apply for the program, would result in still larger MDIs for participants.

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. As shown in Equation 5.5, the variance of the impact estimate is inversely proportional to sample size and, thus, the standard error of the

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.
impact estimate is inversely proportional to the square root of the sample size. This means that the applicant sample would have to be 1.56 times the size of an sample consisting of eligible applicants in order to achieve the same MDI for participants \(N_A = N_E/(.80^2)\).

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 ability of a study with a given sample size to detect meaningful impacts if they exist. In the example, assigning eligible applicants rather than all applicants would entail either a 25 percent increase in MDIs for impacts on participants or a 56 percent increase in the size of the study sample in order to maintain the MDI at established standards of power and significance. 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.

Unfortunately, program staff typically resist suggestions that random assignment be conducted at a relatively late stage of the intake process, after most of those who would drop out or be screened out have been excluded. The reason is that this increases the burden on staff, as they must continue to process those who will ultimately be assigned to the control group as well as those who will be allowed to participate in the program. Moreover, conducting random assignment later in the intake process can raise the sensitivity of staff regarding issues of “fairness” (e.g., risks of raising expectations of some who will be assigned to the control group). Finally, the more contact staff have had with study participants prior to random assignment, the more difficult it tends to be for them to participate in informing individuals of their treatment status. 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. In any
event, it is important to consider where random assignment will be conducted in the “pipeline” and to factor in the costs in terms of lower MDIs and/or larger sample sizes.

**Allocation of the Sample Among Multiple Treatments**

The essence of using experimental designs to estimate the impacts of programs, policies or practices is to be able to compare outcomes for comparable groups of individuals in all respects, except for the program, policy, or practice under study. Thus far, we have not considered issues relevant to deciding the best allocation of the study sample between or among treatment conditions. Below, we first discuss the simple, two-group case: one intervention group and one control group. Then we consider three somewhat more complex situations: (1) the case where a study is designed to estimate the impact of more than one intervention; (2) the case where there are different costs associated with intervention group observations and control group observations; and (3) the case when multiple interventions are implemented in multiple sites.

**Allocating the Sample between the Treatment and Control Groups**

The objective of sample allocation is to maximize the power of the design or, equivalently, to reduce the minimum detectable impact of the study. Thus, in a simple experiment with one treatment group and one control group, we want to choose n_t and n_c such that they yield the smallest possible minimum detectable impact. In a previous section we showed that, for any given allocation of the sample between the intervention and the control group, the minimum detectable impact is inversely proportional to total sample size. Here, we hold total sample size constant and focus on the *allocation* of the sample between the intervention and the control
group—i.e., we pose the problem in terms of choosing the ratio \( n_t/n_c \) for any given total sample. Thus, in the case where there are no practical constraints on the relative size of the two groups, the task is to adopt a random assignment ratio \((n_t/n_c)\) that minimizes the MDI:

\[
MDI = MC \sqrt{\frac{V_t}{n_t} + \frac{V_c}{n_c}}
\]  

(5.13)

It can be shown that the MDI is minimized when \( n_t/n_c = 1 \) (i.e., when equal numbers of individuals are assigned to the intervention and the control group).

In practice there may be other considerations that lead to unbalanced designs—for example, there may be an absolute limit on the size of the treatment group, due to program capacity or it may be necessary to offer a large share of the target population access to the program, policy, or practice under study in order to gain “buy-in” for the study. In these cases, the “constrained” optimal allocation ratio may differ substantially from 1.

When there are multiple interventions, we face a tradeoff among study objectives. Within a fixed total sample size, allocating more of the sample to one intervention will reduce the minimum detectable impact (or increase the power) of the design for estimating the impact of that intervention at the expense of the MDI (or power) of the design to estimate impacts of the other interventions. To determine the optimum allocation in this situation, we must specify the relative importance attached to each of the impact estimates to be derived from the study. This entails specifying an objective function \( MDI_w \) that is a weighted sum of the minimum detectable impacts for the \( k \) interventions being examined in the study:

\[
MDI_w = w_1MDI_1 + w_2MDI_2 + \ldots + w_kMDI_k,
\]  

(5.14)
where \( w_i \) is the “policy weight” attached to the impact estimate for the \( i^{th} \) intervention. Since smaller minimum detectable effects are preferred to larger ones, we wish to allocate the sample to minimize \( \text{MDI}_w \), subject to the constraint:

\[
\sum_{i=1}^{k+1} n_i \leq N \quad (5.15)
\]

where \( n_i \) is the number of individuals randomly assigned to the \( i^{th} \) of \( k+1 \) groups (\( k \) interventions, plus a control group) and \( N \) is the total sample size.

Consider, for example, an experiment with two interventions and a control group. Such an experiment can produce two different impact estimates: the impact of intervention 1 and the impact of intervention 2. If we put equal weight on these two estimates (i.e., if \( w_1 = w_2 = 1 \)), then the objective is to minimize:

\[
\text{MDI}_w = \text{MDI}_1 + \text{MDI}_2 \quad (5.16)
\]

subject to the constraint:

\[
nt_1 + nt_2 + nc \leq N \quad (5.17)
\]

It can be shown that \( \text{MDI}_w \) is minimized (i.e., the power of the design is maximized) when the following condition holds:

\[
\frac{nt_1}{nt_2} = 1 \text{ and } \frac{nc}{nt_1} = \frac{nc}{nt_2} = 2 \quad (5.18)
\]
That is, in the optimal allocation the samples assigned to the two interventions are of equal size and the sample assigned to the control group is *twice* as large as each of the interventions. This means that half the sample should be assigned to the control group and one quarter to each intervention.

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 intervention 1 reduces the minimum detectable effect for that intervention (the first term on the right in equation 5.16), but has no effect on the minimum detectable effect for intervention 2 (the second term); the converse holds for adding an individual to the sample assigned to the second intervention. In contrast, adding an individual to the control group reduces the minimum detectable effect for *both* interventions, because the control group is involved in both experimental comparisons. Of course, since sample size enters into the *denominator* of the MDI formula, 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 intervention 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 generalizes to the case of any number of groups as long as equal weights are placed on each experimental comparison: Sample should be *allocated across groups in proportion to the number of experimental comparisons the group is used in*. If, for example, one study will generate impact estimates for seven different interventions, all using the same control group, each intervention group member will be involved in one impact estimate
and comparison and the control group member 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 intervention and half (7/14) of the sample in the control group.\footnote{The result that half the sample should be allocated to the control group holds only so long as interest lies in the individual estimates of impact. If one is also interested in \textit{comparisons} among the impacts of different interventions, it is no longer optimal to put half the sample into the control group, because comparisons of impacts do not involve the control group; they involve only the treatment groups assigned to the interventions being compared. The more general rule that the sample should be allocated across groups in proportion to the number of experimental comparisons the group is used in \textit{does} hold in this case, however.}

If there is \textit{unequal} policy interest in the different experimental comparisons, the sample allocation must be derived by minimizing the expression for MDI\textsubscript{W} in equation 5.14, with appropriate weights, 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-design evaluation of an ongoing program in which all students who participate in the program within a given time period constitute the study sample. More commonly, however, the binding constraint is not a fixed total universe from which to draw the study sample, but a fixed budget that can be devoted to the interventions being tested and data collection for the study. In this case, minimum detectable impacts may be constrained by the budget rather than potential sample size.

If the cost of assigning an individual sample member to one treatment condition or another (whether an intervention group or a control group) are similar, then having a fixed budget is the same as having a sample size constraint; the total sample size is simply the budget divided by the (uniform) cost of assigning an individual to an experimental group. This is the case, for example, when the interventions under study are supported through funds other than
those supporting the evaluation. This would be the case, for example, in the study of an ongoing program like the Big Brothers-Big Sisters mentoring program (Grossman and Tierney 1998).

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 interventions 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 MRIW (as defined in equation 5.14) subject to the constraint:

\[ n_1c_1 + n_2c_2 + \ldots + n_{k+1}c_{k+1} \leq C \]  

(5.19)

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 (including both the costs of the intervention and data collection), and \( C \) is the total budget for experimental treatment and data collection.

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

\[ n_i / n_c = \sqrt{C_c / C_t} \]  

(5.20)

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 an intensive remedial reading program for low-performing students. Furthermore, suppose that the educational services being studied cost an average of $4,000 per student assigned to the intervention group and nothing for those assigned to the control group, and that data collection costs average of $500 per student in both the intervention and the control groups. Thus, the total cost per student in the intervention group will be $4,500 as compared with $500 for students in the control group. If equal numbers of students are allocated to the intervention and the control groups, the total study budget will support a sample of 100 students in each group. Taking the relative costs of the two groups into account, however, the optimal allocation is to put three times as many students in the control group as in the intervention group. With this allocation, the budget will support 83 students in the intervention group and 250 students in the control group. Minimum detectable impacts under this allocation will be 10 percent smaller than under a balanced design with equal allocation to the treatment and control groups. To achieve this smaller minimum detectable impact with equal-sized groups would have required a 23 percent increase in total sample size, from 200 to 246 and a proportionally larger budget for the study ($615,000 rather than $500,000). Thus, taking account of costs in sample allocation is equivalent to a $115,000 increase in the study budget.
REFERENCES

“Performance Trajectories and Performance Gaps as Achievement Effect-Size
Benchmarks for Educational Interventions.” CITE

Empirical Issues in the Design of Group-Randomized Studies to Measure the Effects of
Interventions for Children, MDRC Working Papers on Research Methodology. New York,
NY: MDRC.


NJ.]

Decker, Paul T., Daniel P. Mayer, and Steven Glazerman (2004). The Effects of Teach For
America on Students: Findings from a National Evaluation. Princeton, NJ:

Garret, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, et al. (2008). The Impact of
Two Professional Development Interventions on Early Reading Instruction and
Achievement. Washington, DC: Institute of Education Sciences, U.S. Department of
Education, NCEE 2008-4030.

403-426.


Exhibit 5.1: Derivation of the Power of the Study Design for True Impact = 5
Exhibit 5.2: Power of the Design for True Impact = 5 with Larger Sample Size

Note to reviewers: In final version, Exhibits 5.1 and 5.2 will be aligned one over the other, so that the change in shape of the sampling distributions will be apparent.
Exhibit 5.3: Calculation of Minimum Detectable Impacts

Sampling distribution under the null hypothesis

Probability

Sampling distribution when true impact = 5

0

\[ 1.64 SE_i + 0.84 SE_i = 2.48 SE_i \]
Exhibit 5.4 Impacts in Natural Units Corresponding to Effect Size of .2, Selected Grades Levels

<table>
<thead>
<tr>
<th>Percent of Average Annual Gains in Reading Test Scores, Mean of Seven Nationally Normed Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades 3-4</td>
</tr>
<tr>
<td>Grades 7-8</td>
</tr>
<tr>
<td>Grades 11-12</td>
</tr>
</tbody>
</table>

Percent of White-Black Performance Gap in Mean Reading Scores, National Assessment of Educational Progress (NAEP)

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
<td>24%</td>
</tr>
<tr>
<td>Grade 8</td>
<td>25%</td>
</tr>
<tr>
<td>Grade 12</td>
<td>30%</td>
</tr>
</tbody>
</table>

Percent of Performance Gap in Mean Reading Scores, Eligible vs Ineligible for Free or Reduced Price Lunch, National Assessment of Educational Progress (NAEP)

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
<td>27%</td>
</tr>
<tr>
<td>Grade 8</td>
<td>30%</td>
</tr>
<tr>
<td>Grade 12</td>
<td>44%</td>
</tr>
</tbody>
</table>

Percent of Performance Gap in Mean Reading Scores, Average School (50th percentile) vs “Weak” School (10th percentile), Two Large Urban School Districts

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td>82%</td>
</tr>
<tr>
<td>Grade 7</td>
<td>111%</td>
</tr>
<tr>
<td>Grade 10</td>
<td>222%</td>
</tr>
</tbody>
</table>

Source: Bloom et al. (2008), Table 3 (average annual gains in reading scores), Table 4 (white-black performance gap and performance gap for eligibles-ineligibles for free or reduced price lunch), Table 6, Districts I and II (performance gap between “weak” and average schools).
Contents

CHAPTER 6: APPLICATIONS OF SAMPLE DESIGN PRINCIPLES .......................................................... 1

TWO-GROUP, ONE-LEVEL DESIGNS .......................................................................................... 3

  Reference Population and Sample Frame .................................................................................. 3
  Units Targeted by the Intervention and Units of Assignment .................................................. 6
  Core Outcome Measures ........................................................................................................... 6
  Minimum Relevant-size Impact (MRI) ....................................................................................... 8
  Background Information Available for the Analysis ................................................................. 9
  Calculating Minimum Sample Size Requirements .................................................................... 10

MORE COMPLEX ONE-LEVEL DESIGNS .............................................................................. 14

  Unbalanced Designs .................................................................................................................. 14
  Sample Attrition ....................................................................................................................... 16
  Tools for Estimating Sample Size Requirements and MDIs in One-level, Two-group Designs ..... 17

TWO-GROUP, TWO-LEVEL DESIGNS .................................................................................... 21

  Accounting for Sample Attrition in Cluster Randomized Designs .......................................... 26
  Tools to Estimate the MD-ES for a Cluster Randomized Controlled Trial ............................. 27

MULTI-GROUP DESIGNS ........................................................................................................... 30

REFERENCES ............................................................................................................................. 33
CHAPTER 6
APPLICATIONS OF SAMPLE DESIGN PRINCIPLES

This chapter illustrates the application of sample design principles to common types of experimental evaluations of educational policies and practices. The ultimate design for any study is determined by a number of factors, many of which are not under the control of the evaluation team. The aim of this chapter is to illustrate how one might efficiently design a study sample to address the core research questions at pre-established standards of evidence, while taking account of programmatic constraints and study contexts.

As discussed in the preceding chapter, there are two standards of evidence that affect decisions regarding sample size and allocation. One is the statistical significance level applied to tests of the null hypothesis of no true impact of the intervention, where the goal is to avoid falsely concluding that there is an impact of the intervention when the observed difference in mean values of the outcome for the intervention and control groups is really due to random sampling error. The other standard, statistical power, is the likelihood that, if the intervention really has an impact equal to or exceeding some pre-specified threshold, the difference between the intervention and control group means will be sufficiently large that it would be judged to provide reliable evidence of a causal impact. For purposes of our examples, we apply the following standards of evidence:

5 percent statistical significance levels for two-tailed tests of the null hypothesis (95 percent confidence intervals around impact estimates)

80 percent power
For those wanting to use alternative levels of statistical significance or power, we provide tools that will allow exploration of the implications of using different standards.

Developing an appropriate sample design for any study requires gathering information that allows reasonable determination of the overall sample size required and how best to allocate the sample between intervention and control conditions. Specifically, we consider the following key factors that impact sample design decisions:

- Reference population and sample frame for the study
- Core outcome measures
- Minimum relevant-size impact for policy and/or practice
- Nature of background information accessible for the analysis
- Units targeted by the intervention and units randomly assigned to intervention or control conditions
- Units for analysis

We begin with a simple example of how to design a sample for a study where the units of assignment to condition are the same as the units for analysis of impacts of the intervention under study. This example entails a study to test a new policy of mandatory summer schools for low-achieving students represented in Exhibit 4.3 above. We then proceed to illustrate how to design samples using other models of random assignment that were discussed in Chapter 4. We provide two types of tools to facilitate the design of simple to moderately complex sample design. Two are sample size estimators that compute, for a particular set of design parameters, the minimum number of units that need to be randomized in order to achieve basic study goals in terms of the minimum size impact that will be detectable at conventional levels of statistical power and precision (see Chapter 5 for discussions of power and precision). One applies to the
case when the units of assignment and analysis are the same; the other pertains to the case where
the units for analysis are clustered within larger units, which are the units that are randomized.
The other two tools are worksheets that allow you to quickly compute minimum detectable
impacts under two basic design scenarios: one in which the unit of random assignment is the
same as the unit of analysis (a one-level design); and the other in which random assignment
occurs at the group level (say teachers) and the analysis is at the individual level (say students
taught by those teachers (a two-level design)

**TWO-GROUP, ONE-LEVEL DESIGNS**

Many policies or practices in education lend themselves to relatively simple two-group
(intervention and control condition), one-level random assignment study designs. For example,
one might want to study the effects of varying the level of feedback students are given on their
homework; measure the benefits of making personal calls to follow up on absent students versus
using automatic call machines to leave messages with parents; examine the impacts on college
matriculation associated with more and less generous student financial aid policies; or see
whether students who have access to computer assisted home work have higher homework
completion rates and attain higher mastery of math concepts than do students with only paper
assignments. We will illustrate how to develop such sample designs using the example of a
study of mandatory summer school for low achieving students (see Exhibit 4.3 above).

**Reference Population and Sample Frame**
In our example of mandatory summer school, the reference population is all U.S. high school
students who are at risk of low achievement. However, practical and cost considerations
generally would result in a much more restrictive sample frame. For example, the sample frame likely would include students from one or a limited number of high schools and, often, a particular cohort of students within the school.

In so far as this is a policy that is not yet widely implemented nor well-researched, a very meaningful study could be conducted under such conditions, recognizing that the applicability of the findings would be limited by the fact that the results are based on a sample drawn from a particular student population and that the intervention and control conditions reflect those prevailing in the communities and schools attended by that population. In addition to the effects of the intervention on the overall sample, there may be interest in being able to estimate impacts for particular subgroups of the reference population, for example, to estimate impacts separately for boys and for girls. If so, this will affect the total sample size needed and it may require that the sample be stratified prior to randomization. In our example, we are going to make the following assumptions:

*The study will be limited to a large urban school district*

*All students entering 9th grade for the first time will be in the sample frame and will be randomly assigned to the intervention or control condition*

*All students in the study sample who are assigned to the intervention condition and subsequently fall into the “low achieving” category during high school will be required to attend summer school*
All students in the study sample who are assigned to the control conditions will experience “business as usual” with respect to summer school referral or attendance regardless of their achievement status.

Other things equal, it generally would be preferable spread the study sample over multiple schools, including schools that differ in size and other characteristics. However, doing so, could add substantially to the cost per sample member of conducting the study. The reason is that, most likely, conducting the study in multiple schools would make it necessary to recruit more sites, train more program staff, and expend more resources collecting data.

For our example, we assume that all of the sample will be drawn from a single school. While this assumption simplifies the sample design, it means that the study results will reflect the effectiveness of the intervention in one particular setting and yield no information regarding the likelihood that the results are sensitive to context. Later, we will illustrate the implications of drawing the sample from a number of different schools and randomizing students within each school (strata) separately.

Although the sample frame for the study is restricted to a single school, it will include all ninth graders in that school. In this way, the study will be able to measure the impacts of the program for all students in that school who fall into the reference population. Students assigned to the intervention condition will be subject to the mandatory summer school program should their achievement fall below the specified threshold at any time during their high school years. We do not restrict the sample to low-achieving 9th graders, because this would omit from the study students who would be subject to the mandatory summer school program for the first time in later years; and we do not include 10th through 12th graders in the study sample because
doing so would result in a sample of students who had different periods of “exposure” to the intervention.

**Units Targeted by the Intervention and Units of Assignment**

In our example, the units targeted by the intervention and the units of random assignment are the same—first-time ninth grade students in the focal high school. In this case, all students in the study sample are subject to the new school environment, whereby any student who exhibits low achievement will be subject to the summer school requirement—that is the intervention. Most likely, only a portion of the students in the school will actually exhibit performance levels sufficiently low as to be subject to the mandate. Judgments about the likely proportion of the sample who will be subject to the summer school requirement will be relevant in estimating the demand for the summer school services, which, in turn, will affect how large the average impact of the intervention would need to be to have practical and/or policy significance (see further discussion of the minimum relevant-size impact, below).

**Core Outcome Measures**

It is critical at the outset of any study to specify the core outcome measures that will be the basis for determining whether or not an intervention achieves its goals. For example, when studying a policy or practice intended to improve overall academic achievement, it makes sense to select a reliable and valid measure of overall academic achievement as the core outcome measure. In contrast, when studying a program aimed at improving mathematics achievement, it makes sense to select a reliable and valid measure of mathematics achievement. In cases where there are
multiple core goals of a program—for example, to improve school attendance and promote better achievement—it makes sense to include both outcomes in the list of core measures. Furthermore, when studying an intervention that might have important unintended adverse effects, it is appropriate to include measures of them among the core outcomes for the purposes of sample design. For example, in a study of “double dose” mathematics instruction, there might be concern that the added time devoted to mathematics would detract sufficiently from reading instructional time that reading achievement might suffer. If this is a real concern, it is important to be able to measure reliably whether or not reading achievement is affected by the intervention.

In selecting core outcome measures, it is important to recognize that there is a trade-off between the number of outcomes measured and the power of the impact analysis. Because the experimental tests of significance must be adjusted for multiple comparisons, for any given sample size, adding more outcomes degrades the power of the experiment (Schochet 2007).

The evaluation team may want to gather information on other (noncore) outcomes for exploratory purposes and possibly to address other research interests. While these other outcome measures are not relevant to critical sample design decisions, it may be desirable to take an a priori look at whether the final sample design will support meaningful analysis of those outcomes, prior to incurring the costs of gathering the data needed only to address impacts on these questions.

In our example, we define two core outcome measures:

(1) Persistence in school, by which we mean whether or not the student remains enrolled in school rather than dropping out of school
(2) Overall academic achievement (measured by a reliable standardized achievement test, such as that used by the district to meet its reporting requirements under No Child Left Behind).

In addition to specifying the outcomes, it is important to gather information regarding the expected means and variances (or standard deviations) of the core outcome measures. This information is necessary to judge the expected size of the sampling error expected under varying sample designs and sizes. It is also useful for gauging what mean levels of the outcomes would be expected under “business as usual” conditions and, more generally, as context for judging the likely relevance for policy or practice of varying size impacts.

**Minimum Relevant-size Impact (MRI)**

As discussed in Chapter 5, it is important to design a study sample so that there is a high likelihood that, should the intervention being studied really have an impact in the meaningfully relevant range (i.e., equal to or larger than the MRI), that fact most likely will be evident in the study findings. In the language introduced in chapter 5, a study sample meeting this condition would have “adequate” statistical power. In order to judge the adequacy of the statistical power for a study with a particular size sample, however, it is first necessary to reach agreement about the minimum relevant impact for policy or practice. What is the smallest size impact that, should it really be the true average impact of the intervention under study, it would be important not to miss detecting it in the study?

Generally, considerable judgment is involved in deciding what the MRI would be for particular outcomes and intervention strategies. These judgments will be affected by both the cost of the intervention, which may or may not be easy to measure, and the value attached to the
changes in the outcomes. Researchers often will be well-advised to consult with key policy
makers and/or education leaders in making such determinations, as policy makers and education
leaders are ultimately an important, and sometimes the main, audience for the research findings.
In any event, they will ultimately decide whether the intervention is sufficiently effective to
warrant adoption. For our example, we select the following MRIs:

For persistence, we set the minimum relevant impact (MRI) at 5 percentage points per
year, where the average persistence rate is 90 percent

For academic achievement, we specify the (MRI) as the equivalent of one-month gain
during the regular school year, where the average achievement gain is 9 months and the
standard deviation of achievement gain is 4.

Background Information Available for the Analysis

Background information can often be used in experimental design studies for two purposes: (1)
to control for any systematic differences between the intervention and control groups; and (2) to
reduce the unexplained variability in outcomes due to factors other than the intervention. The
latter has the effect of decreasing the sampling error in estimates of the impacts of the
intervention and, thus, the sample size required to meet any particular standards of evidence (see
discussion in Chapter 5 of the role of covariates in determining minimum detectable impacts). In
our example we assume the following:

We will be able use to 5 background variables
These variables explain 20 percent of the variance in persistence

These variables explain 50 percent of the variance of academic achievement.

Calculating Minimum Sample Size Requirements

In very simple cases where you have a balanced sample design (i.e., equal size intervention and control group samples) with only one main outcome measure you can apply the simple formula presented in Chapter 5 to determine the minimum sample size requirements to meet evidence standards for each of your main outcomes measures:

\[
N = 4 \times \left( \frac{2.81}{MRI - ES} \right)^2
\]

(6.1)

where,

\(N\) = minimum required sample size (intervention and control group combined)

\(MRI - ES\) = minimum relevant-size impact divided by the standard deviation of the outcome measure

2.81 = a multiplier based on the standards set for the significance level for the null hypothesis test and for statistical power

However, equation 6.1 does not allow you to take account of two features of our illustrative study design parameters: (1) the fact that we have multiple core outcomes and (2) the expectation that we will be able to explain some of the variance in the outcome through use of covariates.
The top two panels of Exhibit 6.1 summarize the key assumptions with regard to standards of evidence and properties of the study sample for our illustrative example. With this information, it is possible to compute how large the sample needs to in order for the study to meet the established evidence standards, taking account of both the fact that there are two outcome measures (albeit in different domains) and that background control variables will be included in the analysis to reduce the unexplained variance in the outcome measures. (Later we will also illustrate how to take account of other common complications, such as unbalanced designs (i.e., designs with other than 50/50 odds of assignment to the treatment or control condition), sample loss at follow-up (attrition), and non-participation in the treatment by those assigned to the intervention group.)

[Insert Exhibit 6.1]

*Background control variables.* When available, baseline (pre-study) measures of the key outcomes, this can reduce the sample size requirements by reducing the unexplained variance in the mean impact estimate. One way to conceptualize the benefits of having background control variables is that they the study sample can be designed to meet a somewhat larger minimum detectable impact than the MRI and the study will still have adequate statistical power to detect the established MRI. Indeed, it is possible to compute an adjusted MRI (AMRI), which accounts for the gains in efficiency through the use of control variables. The Adjusted Minimum Relevant Impact (AMRI) can be expressed as follows:
\[ AMRI = \frac{MRI}{\sqrt{1 - R^2}} \quad (6.2) \]

where,

- \( AMRI \) is the Adjusted Minimum Relevant Impact
- \( MRI \) is the Minimum Relevant Impact
- \( R^2 \) is the variance in the outcome measure explained by the baseline covariates

For example, in the study of a mandatory summer school policy, we anticipated that 20 percent of the variance in achievement test results would be explained by the baseline test score and other control variables. As a result, the AMRI is 11 percent larger than the MRI. Similarly, the AMRI for academic achievement is 41 percent larger than the MRI. As a result, the minimum size sample needed to meet any established evidence standards would be correspondingly smaller.

In balanced sample designs, the addition of covariates can be quite simply incorporated into the formula for minimum required sample size (6.1 above) by substituting the Adjusted Minimum Relevant Impact in effect size units for the MRI-ES:

\[ AMRI - ES_i = \frac{AMRI_i}{SDV_i} \quad (6.3) \]

where,

- \( AMRI - ES \) is the Adjusted Minimum Relevant Impact in effect size units
- \( AMRI \) is the Adjusted Minimum Relevant Impact
- \( SDV \) is the standard deviation
Intuitively, what the AMRI-ES does is convert the MRI from natural units to standard deviation units, which is a common parameter in the minimum required sample size calculation, and then adjusts it upward, when appropriate, to account for the fact that some of the variance in the outcome measure will be explained by the covariates. (Remember that, other things equal, as the MRI (or AMRI) increases, the sample size required to meet any particular set of evidence standards decreases. Being able to explain some of the variance in an outcome measure using covariates means that the AMRI will be larger than your MRI and, therefore, the sample size requirement will be smaller.)

The following variant of the simple formula presented above can be used to compute the minimum required sample size:

\[ N_i = \text{Maximum} \left[ 4 \times \left( \frac{2.81}{AMRI - ES_i} \right)^2 \right] \]  

\[ (6.4) \]

where,

\[ N_i \] is the sample size needed for the study to meet the established evidence standards for the outcomes that has the smallest \textit{AMRI}

\[ N_i \] is the sample size needed for the study to meet the established evidence standards for the outcomes that has the smallest AMRI. Since the AMIR-ES is in the denominator of the right-hand side of equation 6.3, this means that \( N \) is determined by the \textit{maximum} value of that term. In our example, the final minimum sample size requirement will be based on the computation using the AMRI-ES for the measure of persistence in school, as this outcome measure has the smallest
MORE COMPLEX ONE-LEVEL DESIGNS

Rarely are studies designs as simple as the balanced, two-group design used in the examples above. More often than not, it is be necessary or desirable to have unequal size intervention and control groups (commonly referred to as an “unbalanced design”) and, most often, there will be some nontrivial sample loss between through follow-up. Both of these factors complicate the calculation of minimum sample size requirements. For this reason, we have provided a sample size calculator for one-level, two-group designs. With this calculator, it is easy to examine the sensitivity of the sample size requirements to your assumptions about the properties of the study sample, including assumptions regarding unbalanced designs and sample loss.

Unbalanced Designs

For a given total sample size, unbalanced designs are less efficient than balanced designs from the perspective of statistical power. However, it often is convenient (and sometimes necessary) to unbalance the design to address operational concerns. For example, in a study of mandatory summer school, it would not make sense to randomly assign half of the low-performing students to the mandatory summer school requirement if there were only enough summer school programs to accommodate one-fourth of the low performing students. It would be better in this case to adopt one of two alternative strategies: (1) assign half of all students to a non-research group and then randomly assign the remaining sample with equal probabilities to the intervention group.
or the control groups for the study; or (2) unbalance the sample design so that the number of youth randomly assigned to the intervention condition is equal to (or less) than the available number of summer school slots. The “cost” in terms of having an unbalanced sample design (measured by the increase in sample size needed to compensate for the power loss due to the imbalance) is illustrated in Exhibit 6.2.

[Exhibit 6.2]

The following elaboration of formula 6.3 above allows for the possibility of unbalanced sample:

\[
AMRI - ES = 2.81 \star \sqrt{\frac{1 - R^2}{N \star P \star (1 - P) \star r}}
\]

(6.5)

where,

- \( R^2 \) is the variance in the outcome measure explained by the baseline covariates
- \( N \) equals the total number of experimental sample numbers \( (n_T + n_C) \)
- \( P \) equals the proportion of this sample that is randomized treatment, i.e., \( \frac{n_T}{n_T + n_c} \)
- \( r \) = retention rate

When the proportion of the study sample assigned to the intervention condition is between .4 and .6, the penalty for an unbalanced sample design is modest (less than 5 percent). When the assignment ratios are between .3 and .7, the penalty is still less than 20 percent. However, when less than 30 percent of the overall sample is assigned to one of the two conditions, the penalty for an unbalanced sample design gets quite high. For example, as seen in the last row of Exhibit 6.2,
whereas it requires only a 20 percent larger sample to compensate for moving from a 50/50 allocation to a 30/70 allocation, it requires a 56 percent larger sample to compensate for a 20/80 allocation and 278 percent larger sample to compensate for a 10/90 allocation.

**Sample Attrition**

Some loss of sample between random assignment and follow-up is common in intervention studies. By “loss of sample,” we mean that the researcher is unable to obtain follow-up data for some of the study sample; it does not refer to the sample members assigned to the intervention condition who for one reason or another chose not to participate or to drop out of the treatment prematurely. (The latter is a different problem that is discussed in a separate section below.)

Sample attrition may derive from multiple sources. For example, students may leave school altogether, they may change schools, or they may miss school on the date when tests or surveys are administered. In some cases, limited project resources may lead to a decision to follow up only a subset of the study sample or the project schedule may mean that some sample members (for example late enrollees) are not eligible for particular waves of follow-up data collection. In designing a sample, it is important to estimate the expected sample loss for the core outcomes and to take account of this in determining how large the initial study sample needs to be. In general, sample attrition can be accounted for quite simply by increasing the expected sample size requirements by the inverse of the expected sample attrition rate. However, it is important to note that, as a rule of thumb, the study implementation plan should yield an 80 percent or better response to follow up (at most a 20 percent sample loss) and there should not be substantially different rates of sample loss between the intervention and control groups. Both
high and differential attrition rates can adversely affect the internal validity of the impact estimates, even from an experimental design study.

Tools for Estimating Sample Size Requirements and MDIs in One-level, Two-group Designs

We have provided two tools to assist in the development of two-group, one-level sample designs. One is a Minimum Required Sample Size Calculator and the other is a minimum detectible impact (MDI) calculator. These complement readily accessible power calculators that are available through Optimal Design (Raudenbush et al. 2004) and elsewhere.

The Minimum Required Sample Size Calculator (Tool 6.1). This tool allows the user to plug in evidence standards and study characteristics and, through a simple iterative process, estimate the minimum sample size required. The first time you use the tool, it will contain the study standards and sample information for the illustrative case presented in Exhibit 6.1. However, the tool allows you to over-write any of these assumptions. You can use the data in Exhibit 2 to guide the selection of starting values for the iterative procedure that generates sample size estimates. So, for example, if you have an MRI of .5 and are planning on a balanced sample, you might start your sample estimation by inputting a starting sample size of 126, whereas if your MRI is .1 and you are expecting to have an unbalanced sample design with about 20 percent of the sample assigned to the intervention group, you would start with a sample value of 4935.

In almost all cases, the worksheet converges on the minimum sample size estimate within three iterations. The calculator requires the user to input the following 7 parameters:
1. Significance level for the null hypothesis test (usually 5 percent)

2. One-or two-tailed test of the null hypothesis (usually 2-tailed)

3. Statistical power standard (usually 80 percent)

4. Minimum relevant impact expressed in standard deviation units (MRI-ES), based on judgments of the policy and/or practice community

5. Proportion of the sample allocated to the intervention group

6. Proportion of the variance in the outcome measure that can be explained by background control variables

7. Sample retention at follow-up

In addition, the user must input a “guesstimate” of the minimum required sample size (row 8) and continue to revise this estimate with the program generated output estimate, until the two numbers converge. The specific formula used in the calculation is as follows:

\[
N = \frac{1 - R^2}{P(1 - P)r} \left( \frac{M_{\hat{N} - k^* - 2}}{MRI - ES} \right)^2
\]

(6.6)

where,

\( N \) is the Minimum Required Sample Size

\( P \) equals the proportion of this sample that is randomized treatment, i.e, \( n_T / (n_T + n_C) \)

\( r \) = retention rate

\( R^2 \) is defined as the proportion of pooled unexplained variation in the outcome within experimental groups predicted by covariates

\( \hat{N} \) is the pre-estimated sample size

\( k^* \) equals the number of covariates
Multiplier for one-tailed test: \( M_{\hat{N} - k^* - 2} = t_\alpha + t_{1-\beta} \) with \( \hat{N} - k^* - 2 \) degrees of freedom

Multiplier for two-tailed test: \( M_{\hat{N} - k^* - 2} = \frac{t_\alpha}{2} + t_{1-\beta} \) with \( \hat{N} - k^* - 2 \) degrees of freedom

\( \alpha \) is the type-I error

\( \beta \) is the type-II error, i.e., \( 1 - \beta \) is the power

It typically requires 2 to 4 iterations of the calculator to reach the final \( N \). To begin the estimation process, the user must provide starting values of all parameters, including a “guestimate” of the sample size, \( \hat{N} \). If the returned sample size estimate differs from the “guestimate,” the user uses the returned sample size estimate as \( \hat{N} \) in the next iteration. This process continues until the estimate of \( N \) returned by the calculator equals \( \hat{N} \).

*Minimum Detectable Effect Calculator (Tool 6.2).* This tool can be used to estimate the minimum size impact that will be detectable (MDI) at specified levels of significance and power, given the properties of the study sample. This is most useful for making quick calculations to determine whether a particular proposed sample design or a particular analysis sample could detect with impacts equal to or smaller than those judged to have policy or practical relevance—the MRI. However, it also useful when evaluating the standards of evidence for published research studies.

This calculator requires that the user input 9 pieces of data. The first three are the standards of evidence required for the sample size estimator; the other six are as follows:

1. Standard deviation of the outcome measure
2. Proportion of the sample allocated to the intervention group
3. Proportion of the variance in the outcome measure that can be explained by background control variables

4. Sample retention at follow-up

5. Number of control variables included in the analysis (relevant mainly for small samples)

6. Total sample size

With this information, the worksheet outputs an estimate of the minimum detectable impact expressed in natural units (the MDI) and the minimum detectable impact expressed in standard deviation units (the MDI-ES). Specifically, the worksheets uses a formula adapted from Bloom (2006, page 12):

\[ \text{MDI} - \text{ES} = M \sqrt{\frac{1 - R^2}{N \cdot P \cdot (1 - P) \cdot r}} \]

(6.7)

where,

- \( N \) equals the total number of experimental sample numbers \((n_T + n_C)\),
- \( P \) equals the proportion of this sample that is randomized treatment, i.e., \( n_T / (n_T + n_C) \)
- \( r \) = retention rate
- \( k^* \) equals the number of covariates
- \( \sigma^2 \) equals the pooled outcome variance across subjects within experimental groups
- \( R^2 \) is defined as the proportion of pooled unexplained variation in the outcome within experimental groups predicted by covariates

Multiplier for one-tailed test: \( M_{N-k^*-2} = t_{\alpha} + t_{1-\beta} \) with \( N-k^*-2 \) degrees of freedom
Multiplier for two-tailed test: $M_{N-k^*-2} = t_{\alpha/2} + t_{1-\beta}$ with $N-k^*-2$ degrees of freedom

In cases where the MDI is equal to or less than the MRI, the study sample would be judged to have adequate power to meet the specified evidence standards. In cases where the MDI is larger than the MRI, the study would be judged to be under-powered.

This tool is set up in a way that allows the user to compute MDIs for three scenarios. For example, it might be useful to be able to compare the MDIs for different outcomes within a study, to compute MDIs under different assumptions about the sample characteristics, and/or to compute MDIs under different standards of evidence. When you first open the worksheet, it will contain three sets of MDI estimates that differ as a result of different assumptions about the sample size.

**Two-group, two-level Designs**

Two circumstances could lead one to recommend using what is referred to as a cluster randomized design. One is the case in which the program, policy or practice under study is designed to affect entire groups of individuals. For example, peer assisted learning is directed at small groups of students within a classroom; curricula changes would be directed at affecting instruction of whole classrooms, grades, and/or schools; and distributed leadership training is intended to affect whole schools. In such cases, one would expect the fidelity of implementation to be jeopardized if it were tested in a setting other than the natural group unit. In other cases, it may be possible to deliver the intervention with fidelity in settings where individuals are assigned to the intervention or control condition, but there is reason to believe that in doing so,
there may be spillover effects on the control (business as usual) group. For example, if one were
to test a model of teacher professional development in mathematics education using random
assignment of teachers within a school, it would be reasonable to worry about the possibility that
control group teachers might, in fact, reap some indirect benefits from the professional
development as a result of interaction with peer teachers.

In cases where there are substantial risks of spillover or where the intervention simply
requires implementation at the group level, the study design should rely on cluster
randomization. In this case, the total size of the analysis sample for outcomes measured at the
level of the units within clusters (e.g., students clustered in classrooms or teachers clustered in
schools) will be larger than under a design where these individual sample units were randomized.
This is the case in the example of a study of the benefits of providing middle school teachers
with access to and training in educational technology (see Exhibit 4.4 above). In this example,
eligible teachers are randomly assigned to the intervention condition or to a business as usual
control condition. If the only outcomes of interest were those measured at the teacher level, one
would apply the methods described in the previous section for two-group, one-level random
assignment designs to decide on the optimal sample size and allocation. However, when the
primary outcomes of interest pertain to the students who are nested within the classrooms of the
teachers who are randomly assigned to the intervention or control condition, the sample design is
more complicated.

As discussed in Chapter 5, the total sample size required to achieve evidence standards in
a cluster randomized trial depends on several additional assumptions about the properties of the
study sample. These include the following:
1. Average number of units per cluster \( (N_C = \text{analysis units per cluster, assumed to be 20 in our example}) \)

2. Clusters are allocated equally between the intervention and the control condition \( (P_T \text{ assumed to be .5 in our example}) \)

3. Intra-class correlation \( (\text{ICC} = \text{proportion of the total variance in the outcome measure that is between clusters, assumed to be .1 in our example}) \)

4. Proportion of the within cluster variance in the outcome measure that is explained by background control variables \( (R_t^2 \text{ assumed to be .5 in our example}) \)

5. Proportion of the between cluster variance in the outcome measure that explained by background control variables \( (R_c^2, \text{ assumed to be .2 in our example}) \)

6. Number of group level background control variables at the cluster level used in the analysis \( (\text{assumed to be 5 in our example}) \)

7. Sample retention rate for clusters and for individuals within clusters \( (\text{assumed to be 100 percent in our example}) \)

**Estimating Sample Size and MDIs for Two-level Cluster Random Assignment Designs**

Tool 6.3 estimates the sample size needed for a study to meet established evidence standards in cases where the unit of analysis is individuals (for example, students or teachers) who were randomly assigned in clusters (for example, whole classes of students or schools of teachers). As with the one-level sample size estimator described above, this one also works through a simple iterative process, whereby you input a “guestimate” of the minimum number of clusters that need to be randomized in order to meet the evidence standards, compare the resulting minimum estimate of the number of clusters with your guestimate and repeatedly replace the “guestimate”
with the most recent “estimate” until the two numbers converge. Convergence typically occurs within two or three iterations.

Exhibit 6.3 provides a quick guide to approximate cluster sample size requirements under various basic assumptions that can provide starting values. As the intra-class correlation (proportion of the variance in the outcome that is between rather than within clusters) approaches zero, the minimum sample size approximates that for a simple random sample and the minimum number of cluster assigned to intervention or control condition will approximate the total size of the analysis sample divided by the number of units per cluster. The figures in Exhibit 6.3 assume there are 20 observations per cluster. Other things equal, the sample size needed to meet any particular evidence standard will be higher for samples with higher intra-class correlations. In the simple case where there are no covariates, for example, the minimum sample size required to meet the evidence standards for our illustrative design of a sample for the study of access to and professional development in use of technology products would require a sample of only 42 classrooms, assuming that teachers and students were randomly assigned to classrooms and students (the ICC = 0); it would be increases by 115 classrooms, assuming the ICC is .10; 190 if the ICC is .2; and 268 if the ICC = .3.

An important take-away message from Exhibit 6.3 is that it is very important to obtain as good information as possible regarding the likely size of the ICC prior to finalizing a study sample design. This message is reinforced by the empirical estimates of ICCs for various samples and outcomes compiled by Hedges and Hedberg (2007) and Schochet (2005).
Importantly, to the extent that there is reason to expect that background control variables will be available for the analysis, it is useful to incorporate this information into the sample size estimates. In so far as there are two components to the overall variance in the outcome estimates that affect the standard error of the impact estimate—one associated with the units of analysis (students in our example) and the other associated with the level of assignment to intervention condition (teachers, in our example)—it is necessary to specify assumptions about the expected proportion of each variance component that will be explained by covariates and incorporate this information into the sample size calculation.

Exhibit 6.4 shows the sensitivity of the minimum sample size requirements to variations in the assumed predictive power of background control variables you expect to be able to include in your analysis. The top panel assumes the analysis will include only covariates specific to the units of analysis (e.g., students) and not the clusters of assignment (e.g., teachers). When the ICC = 0, the sample size is reduced in proportion to the portion of the total variance in the outcome that is explained by the control variables. However, as the ICC moves away from zero, having good estimates of the proportions of the variances in outcomes between clusters ($R_c^2$) and within clusters ($R_t^2$) that will be explained by covariates will be important in getting good estimates of the minimum size sample you need to ensure your study findings will meet established evidence standards.

[Insert Exhibit 6.4]

For our illustrative two-level, two group study design, we provide a “benchmark estimate” of the sample size based on the assumptions stated above, as well estimates for two
alternative scenarios: (1) an optimistic scenario in which the ICC is assumed to be zero; and (2) a more pessimistic scenario in which there are no efficiency gains from including background control variables in the analysis ($R_i^2 = R_c^2 = 0$). Exhibit 6.4 details the key assumptions underlying the benchmark sample size estimate, as well as the sample size estimates for the two alternative scenarios.

To meet evidence standards under the benchmark assumptions requires a study sample of 84 teachers randomly assigned to the intervention or control condition and a total of 1,660 students. If the reality is that none of the variance in the outcome measure was explained by control variables, it would be necessary to have a sample of 115 teachers and 2,290 students. However, if it is that none of the variance in the outcome measure was between teachers, but instead all was within the classroom, the evidence standards could be met with a sample of only 20 teachers randomly assigned to the intervention or control condition and the approximately 400 students in these teachers’ classrooms. The latter is the result one would come up with using the one-level, two group sample size estimator (Tool 6.1 above).

**Accounting for Sample Attrition in Cluster Randomized Designs**

As with the case of one-group, two-level designs, discussed above, it is equally important in two-level designs to expect and account for normal levels of sample attrition. Tool 6.3 includes two sample retention parameters—one for clusters and one for the analysis observations within the cluster. In essence, the sample size calculator adjusts the sample minimum sample size estimates upward to account for the expected sample loss. In essence, the estimator “inflates” the sample size estimates for clusters and for analysis units by the inverse of the sample retention rate.
As was discussed in Chapter 5, one should not assume that the only implication of sample loss is the need to have a larger study sample to begin with. The more serious issue with sample attrition relates to the potential that sample loss is nonrandom and, possibly, differential between the intervention and control groups. Similar levels and types of nonrandom sample loss among the intervention and control groups limits the generalizability of the study findings, even to the original sample frame for the study; differential sample loss between the intervention and control groups poses serious threats to the internal validity of the study.

**Tools to Estimate the MD-ES for a Cluster Randomized Controlled Trial**

Similar to the MDI calculator for two-group, one-level study designs, Tool 6.4 is an MDI calculator for cluster randomized controlled trials. This tool computes the MDI under specified standards of evidence and study characteristics. For example, following Bloom (2007, page 17), and adding parameters to account for sample retention, the calculator computes the MD-ES for a one-stage cluster sample as follows:

\[
\text{MDES}(\tilde{\beta}_0) = M_{j-g' \geq -2} \sqrt{\frac{\rho(1-R_g^2) + (1-\rho)(1-R_{g'}^2)}{P(1-P)r_JJ}}
\]

(6.8)

Where the multiplier for one-tailed test is \( M_{j-g' \geq -2} = t_\alpha + t_{1-\beta} \) with \( J-g*\)-2 degrees of freedom and that for a two-tailed test is \( M_{j-g' \geq -2} = t_{\alpha/2} + t_{1-\beta} \) with \( J-g*\)-2 degrees of freedom, and:

- \( J = \) the total number of clusters
- \( \rho = \) intra-class coefficient (ICC) = \( \frac{\tau^2}{\tau^2 + \sigma^2} \)
- \( \tau^2 = \) between group-level variance (unconditional)
\[ \sigma^2 = \text{individual-level variance (unconditional)} \]

\[ R_1^2 = \text{the proportion of individual variance (at level one) predicted by covariates (see Appendix 6.1)} \]

\[ R_2^2 = \text{the proportion of group variance (at level two) predicted by covariates (see Appendix 6.1)} \]

\[ g^* = \text{the number of group covariates used (n.b.: the number of individual covariates does not affect the number of degrees of freedom)} \]

\[ r_1 = \text{sample retention rate for level 1 (students)} \]

\[ r_2 = \text{sample retention rate for level 2 (clusters)} \]

It is especially useful in helping determine whether studies were, in fact, designed so as to be able to reliably detect impacts of a meaningful size should they exist. As noted above, it is especially useful to be able to easily estimate the minimum detectable impacts for studies that show no evidence of statistically significant impacts for the interventions under study. The reason is that the conclusion one draws from such studies is quite different depending on the size of the MDI and its relationship to the minimum size impact that would be relevant for policy or practice.

Using the *Minimum Detectable Effect Calculator for Cluster Randomized Controlled Trials* (Tool 6.4) requires the same information about standards of evidence and the property of the study sample required to determine the minimum sample size requirements. In this case, however, the user inputs the sample size information and the tool computes the minimum detectable effect in both standard deviation units and in natural units: The number of clusters and individual analysis units in the study (N_{Retained} and N_{CRetained}); the proportion of the sample allocated to the intervention condition (P_{TRetained Sample}); N_c (the average number of analysis units...
per cluster at follow-up); the ICC; the proportion of the within and between cluster variance that is explained by covariates (R_1^2 and R_c^2); and the number of cluster-level background control variables used in the analysis (which is important only for small sample sizes). In addition, in order to be able to compute the MDI in natural units, it is necessary to know the standard deviation of the outcome measure.

When the reader first opens up Tool 6.4, you will find input parameters describing the evidence standards and study sample for our illustrative two-group, two-level study in the first column, labeled “version 1.” However, the tool also provides two additional input/output columns that allow the user to experiment with variations in the standards of evidence or the assumptions about the study characteristics. When the user first opens the tool, versions 2 differs from version 1 in terms of the variance explained by the control variables measured at the analysis (individual) level. The user is free to change any of the standards of evidence or assumptions about the properties of the study sample and see how these affect the MDI.

Exhibit 6.5 presents summarizes provides illustrates the use of the MDE-ES calculator using sample design information for [X] recently completed cluster randomized controlled trials released by the Institute of Education Sciences. What is notable from this table is the high degree of variability in the MDIs.

Review note: Once the table is completed, we will add a few sentences describing the variability and where likely is and is not planned.

[Exhibit 6.5]
MULTI-GROUP DESIGNS

Sometimes there are reasons to conduct a “multi-arm” trial, meaning that there are multiple intervention conditions being tested simultaneously. As discussed in Chapter 5, assuming that you are equally interested in estimating impacts for each of the interventions, the optimal sample design will entail placing half of the study sample in the control group and half in the treatment group, resulting in an unbalanced study design estimating impacts of each intervention. Then, you would proceed to use the appropriate sample size estimator tool to determine the minimum number of units that need to be allocated to the intervention and control condition in order to meet established evidence standards.

For example, Exhibit 4.8 above profiled a model of random assignment of teachers to one of three treatment conditions—(1) treatment condition 1, which consists of access to educational technology and ongoing training and professional support to promote effective use of the technology; (2) treatment condition 2, which consists of access to educational technology and one-shot training in its use; or (3) business as usual. (This is a simplified version of the design for the recently completed randomized controlled trial of educational technology products conducted by the Institute of Education Sciences (Dynarski et al. 2007). In a situation like this, assuming there is equal interest in the impact estimates of each intervention; one would first establish the optimal ratio of teachers to be assigned to each intervention condition, relative to the control condition:

\[ P_T = \frac{1}{2 \times \# \text{intervention treatment groups being assessed}} \]

or .25, in our example.
With this parameter set, you can proceed to use tool 6.3—the sample size estimator for two-level, two group designs to determine how large the control group and each treatment group needs to be in order to meet the evidence standards. The total sample, in our example, will be four times the size of each treatment group. If there were three treatment groups, \( P_T \) would = \( .167 \), and so on.

Exhibit 6.6 illustrates the sample size requirements under three scenarios. In essence, there are economies of scale in conducting tests of multiple intervention alternatives simultaneously; there is a “sale” price in terms of sample units, because the control group observations are used multiple times. This multiple use of control group observations offsets the inefficiencies associated with an unbalanced design.

[Insert Exhibit 6.6]

A similar logic would apply to situations of a multi-arm, one-level randomized controlled trial. In this case, the only difference is that one would determine the initial sample allocation using Tool 6.1, not 6.3.
Appendix 6.1

Calculating $R_1^2$ and $R_2^2$

Unconditional Model: $Y_{ij} = \alpha + e_j + \varepsilon_{ij}$

$\varepsilon_{ij} \sim N(0, \sigma_0^2)$  \hspace{1cm} $e_j \sim N(0, \tau_0^2)$  \hspace{1cm} $\rho_0 = \frac{\tau_0^2}{\tau_0^2 + \sigma_0^2}$

Conditional Model: $Y_{ij} = \alpha + \beta_1 W_j + \beta_2 X_{ij} + e_j + \varepsilon_{ij}$

$\varepsilon_{ij} \sim N(0, \sigma_1^2)$  \hspace{1cm} $e_j \sim N(0, \tau_1^2)$  \hspace{1cm} $\rho_1 = \frac{\tau_1^2}{\tau_1^2 + \sigma_1^2}$

$R_1^2 = \frac{\sigma_0^2 - \sigma_1^2}{\sigma_0^2}$  \hspace{1cm} $R_2^2 = \frac{\tau_0^2 - \tau_1^2}{\tau_0^2}$  \hspace{1cm} $R_{total}^2 = \frac{(\tau_0^2 + \sigma_0^2) - (\tau_1^2 + \sigma_1^2)}{\tau_0^2 + \sigma_0^2}$
References


Garret, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, et al. (2008). The Impact of Two Professional Development Interventions on Early Reading Instruction and


http://www.gse.upenn.edu/pimfer/docs/RAM%20Designing%20Studies%20Adequate%20Stats%202011.13.06.pdf


Exhibit 6.1: Assumptions for Simple Random Assignment Studies Estimate the Effects of Mandatory Summer School for Low Achieving High School Students

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Persistence</th>
<th>Academic Achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standards of Evidence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance level of the null hypothesis test</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>(probability of concluding there is a nonzero impact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>when there is not)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power (probability of obtaining a statistically</td>
<td>80%</td>
<td>80%</td>
</tr>
<tr>
<td>significant impact estimate when the true impact is</td>
<td></td>
<td></td>
</tr>
<tr>
<td>at least as great as the minimum relevant impact (MRI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum Relevant Impact (MRI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum Relevant Impact in natural units (MRI&lt;sub&gt;natural units&lt;/sub&gt;)</td>
<td>5.00%</td>
<td>1.00</td>
</tr>
<tr>
<td>MRI expressed in standard deviations (MRI-ES)</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Properties of the Study Sample</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of the sample assigned to treatment</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>condition (N&lt;sub&gt;t&lt;/sub&gt;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean of the outcome measure for &quot;business as usual</td>
<td>90%</td>
<td>9.00</td>
</tr>
<tr>
<td>condition&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation (Sdv) of the outcome measure for</td>
<td>30%</td>
<td>4.00</td>
</tr>
<tr>
<td>the business as usual condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of background control variables ti be included</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>in the analysis (N&lt;sub&gt;v&lt;/sub&gt;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated proportion of the variance in the outcome</td>
<td>20%</td>
<td>50%</td>
</tr>
<tr>
<td>measure explained by the background control variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>explained by the background control variables (R&lt;sup&gt;2&lt;/sup&gt;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Adjusted Minimum Relevant Impact (AMRI)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Minimum Relevant Impact in natural units (AMRI)</td>
<td>5.59%</td>
<td>1.41</td>
</tr>
<tr>
<td>AMRI expressed in standard deviations (AMRI-ES)</td>
<td>0.19</td>
<td>0.35</td>
</tr>
</tbody>
</table>

| Minimum Required Sample Size to Meet Evidence Standards | 910 | 253 |

Note: The minimum required sample size is based on the formula in Tool 6.1.
### Exhibit 6.2: Illustrative Sample Size Requirements to Meet Commonly Accepted Standards of Evidence Requirements (80 Percent Power and 5 Percent Significance Level for the Null Hypothesis Test)

| Minimum relevant impact in standard deviation units (MRI-ES or AMRI-ES) | Proportion of the study sample assigned to the treatment group (p_t) |
|---|---|---|---|---|
| 0.1 | 0.5 | 3158 | 3290 | 3760 | 4935 | 8773 |
| 0.2 | 0.4 | 790 | 823 | 940 | 1234 | 2193 |
| 0.3 | 0.3 | 351 | 366 | 418 | 548 | 975 |
| 0.4 | 0.2 | 197 | 206 | 235 | 308 | 548 |
| 0.5 | 0.1 | 126 | 132 | 150 | 197 | 351 |
| 0.6 | 0.0 | 88 | 91 | 104 | 137 | 244 |
| 0.7 | 0.0 | 64 | 67 | 77 | 101 | 179 |
| 0.8 | 0.0 | 49 | 51 | 59 | 77 | 137 |
| 0.9 | 0.0 | 39 | 41 | 46 | 61 | 108 |
| 1.0 | 0.0 | 32 | 33 | 38 | 49 | 88 |

**Penalty for Unbalanced Design**

\[ \left( \frac{N_{\text{unbalanced}}}{N_{\text{balanced}}} \right) \]

1.00 1.04 1.19 1.56 2.78

Note: The minimum sample size is that required to be able to detect true impacts as small as that designated as the minimum relevant impact with 80 percent power and 95 percent confidence in situations where the study sample is randomly assigned to treatment or control condition. The formula for computing the minimum sample size required is the following:

\[ N = N_t + N_c = \left( \frac{2.81}{\text{MRI-ES}} \right)^2 / \left( \frac{N_t}{N} \left( 1 - \frac{N_t}{N} \right) \right) \]

where

- \( N \) = total sample size for analysis
- \( N_t \) and \( N_c \) = the number assigned to the treatment and control groups, respectively
- 2.81 represents a multiplier that takes account of the statistical significance standard for the null hypothesis test (set at .05, two-tailed test) and the power level (80%) in this example.
- MRI-ES = the minimum relevant impact in standard deviation units the study needs to be able to detect at the designated standards of evidence (statistical power and precision)
Exhibit 6.3: Illustrative Minimum Number of Clusters Randomized to Achieve MDE-ES of .2 (.05 significance; 80% power; and 20 units per cluster), by variance explained by covariates and ICC

<table>
<thead>
<tr>
<th>Percent of Variance Explained</th>
<th>0.00</th>
<th>0.10</th>
<th>0.20</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only individual-level covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.3</td>
<td>265</td>
<td>262</td>
<td>259</td>
<td>256</td>
<td>254</td>
<td>251</td>
<td>248</td>
</tr>
<tr>
<td>ICC = 0.2</td>
<td>190</td>
<td>187</td>
<td>183</td>
<td>180</td>
<td>177</td>
<td>174</td>
<td>171</td>
</tr>
<tr>
<td>ICC = 0.1</td>
<td>116</td>
<td>111</td>
<td>107</td>
<td>104</td>
<td>100</td>
<td>97</td>
<td>93</td>
</tr>
<tr>
<td>ICC = 0.0</td>
<td>40</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>24</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Only cluster-level covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.3</td>
<td>265</td>
<td>241</td>
<td>217</td>
<td>194</td>
<td>170</td>
<td>146</td>
<td>123</td>
</tr>
<tr>
<td>ICC = 0.2</td>
<td>190</td>
<td>174</td>
<td>158</td>
<td>142</td>
<td>126</td>
<td>111</td>
<td>95</td>
</tr>
<tr>
<td>ICC = 0.1</td>
<td>115</td>
<td>107</td>
<td>99</td>
<td>91</td>
<td>83</td>
<td>75</td>
<td>67</td>
</tr>
<tr>
<td>ICC = 0.0</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Individual and cluster-level covariates, similar predictive power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.3</td>
<td>265</td>
<td>238</td>
<td>212</td>
<td>185</td>
<td>159</td>
<td>132</td>
<td>106</td>
</tr>
<tr>
<td>ICC = 0.2</td>
<td>190</td>
<td>171</td>
<td>152</td>
<td>133</td>
<td>114</td>
<td>95</td>
<td>76</td>
</tr>
<tr>
<td>ICC = 0.1</td>
<td>115</td>
<td>103</td>
<td>92</td>
<td>80</td>
<td>69</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td>ICC = 0.0</td>
<td>40</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>24</td>
<td>20</td>
<td>16</td>
</tr>
</tbody>
</table>

Note 1: These estimates are based on the formulas used in Tool 6.2. They pertain to cases where the units of assignment are the same as the units of analysis (e.g., analysis of student outcomes when students were randomly assigned to intervention or control condition or analysis of teacher outcomes when teachers were randomized.

Note 2: Other things equal, the number of sample clusters needed to achieve a larger (smaller) MDE-ES will can be approximated by multiplying the above numbers by the inverse of the $(.2/MDE-ES^*)$, where MDE-ES* is the alternative minimum detectable effect size. For example, an MDE-ES of .1 would require samples approximately 4 times as large as those reported in this table, while an MDE-ES of .3 would require a sample .44 times as large as those reported in this table.
Exhibit 6.4: Illustrative Sample Size Estimates for Two-Group, Two-Level Designs

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Benchmark Assumptions</th>
<th>Alternative Scenario 1</th>
<th>Alternative Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standards of evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Significance level for null hypothesis test (generally 5%)</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>2. Two-tailed or one-tailed Test? (generally set at 2)</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3. Statistical Power Standard (generally 80%)</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>4. MRI-ES (Minimum relevant impact in standard deviation units)</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Properties of the study sample</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Nc (average number of analysis units per cluster)</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>6. Intraclass correlation (ICC or the proportion of the variance in the outcome that is between clusters)</td>
<td>0.10</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>7. Pr (proportion of the sample allocated to the intervention condition)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>8. R² (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)</td>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>9. R² (proportion of the variance in the outcome explained by background control variables used in the cluster-level models)</td>
<td>0.20</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10. Number of group level background control variables used</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11. Sample retention for the analysis sample at follow-up (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>12. Sample retention rate for the cluster level observations (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>14. Estimated Minimum Required Number of Clusters: If not equal to guestimated value in previous row, replace the guestimate with the value in this row.</td>
<td>83</td>
<td>115</td>
<td>20</td>
</tr>
<tr>
<td>15. Estimated Minimum Required Number of Analysis Units: If not equal to guestimated value in previous row, replace the guestimate with the value in this row.</td>
<td>1651</td>
<td>2292</td>
<td>395</td>
</tr>
</tbody>
</table>


*For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).*
Exhibit 6.5: Study Characteristics and Minimum Detectable Impacts for Cluster Randomized Trials Recently Released by IES

Review Note: This table needs to be completed. The "sample case" will disappear; only real cases will remain. Full references to the studies will be provided, including the web links.

<table>
<thead>
<tr>
<th>Study and Outcome Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Case</td>
</tr>
<tr>
<td>Reading test (grade 1)</td>
</tr>
<tr>
<td>Teacher Retention</td>
</tr>
<tr>
<td>Reading test scores</td>
</tr>
<tr>
<td>Math test scores</td>
</tr>
<tr>
<td>Reading/Treatment</td>
</tr>
<tr>
<td>Reading/Pooled treatments</td>
</tr>
<tr>
<td>Study Characteristics</td>
</tr>
<tr>
<td>N CLUSTERS</td>
</tr>
<tr>
<td>PT</td>
</tr>
<tr>
<td>NIC</td>
</tr>
<tr>
<td>ICC</td>
</tr>
<tr>
<td>R²</td>
</tr>
<tr>
<td>R² s</td>
</tr>
<tr>
<td>N CVar</td>
</tr>
<tr>
<td>SDV-Outcome</td>
</tr>
<tr>
<td>Minimum Detectable Impacts</td>
</tr>
<tr>
<td>MDI-ES</td>
</tr>
<tr>
<td>MDI- Natural</td>
</tr>
<tr>
<td>Units</td>
</tr>
</tbody>
</table>

Note: The MDI-ES calculations in this worksheet are based on the formula used in Tools 6.2 and 6.4, depending whether or not the outcome pertains to the units that were randomly assigned or to subunits within the assignment clusters.

Information on the various studies can be found in the following documents:
1. Dynarski et al. (2007)...
2. Glazerman et al. 2006...teacher induction
3. PD Math TWG 2007
4. AIR 2004...PD math NOTE: 3-group design.
5. James-Burdamy et al. 2006. Reading comp per intervention
6. James-Burdamy et al. 2006. Reading comp per intervention/ pooled across 4 interventions

Made up data for now. This table will be completed using data from the study reports.
### Properties of the study sample

<table>
<thead>
<tr>
<th></th>
<th>One Intervention Group</th>
<th>Two Intervention Groups</th>
<th>Three Intervention Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of treatment conditions</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2. Nc (average number of analysis units per cluster)</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3. Intraclass correlation (ICC or the proportion of the variance in the outcome that is between clusters)</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>4. P&lt;sub&gt;T&lt;/sub&gt; (proportion of the sample allocated to the intervention condition)</td>
<td>0.50</td>
<td>0.33</td>
<td>0.25</td>
</tr>
<tr>
<td>5. R&lt;sup&gt;2&lt;/sup&gt; (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>6. Number of group level background control variables used</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>7. Sample retention for the analysis sample at follow-up (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>8. Sample retention rate for the cluster level observations (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Estimated Minimum Sample Sizes

<table>
<thead>
<tr>
<th></th>
<th>One Intervention Group</th>
<th>Two Intervention Groups</th>
<th>Three Intervention Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Clusters for Each Treatment/Control Comparison</td>
<td>83</td>
<td>91</td>
<td>108</td>
</tr>
<tr>
<td>14. Clusters Total Across All Treatment and Control Groups</td>
<td>83</td>
<td>122</td>
<td>162</td>
</tr>
<tr>
<td>15. Analysis Units for Each Treatment Control Comparison</td>
<td>1651</td>
<td>1823</td>
<td>2160</td>
</tr>
<tr>
<td>16. Analysis Units for All Treatment-Control Comparisons</td>
<td>1651</td>
<td>2430</td>
<td>3241</td>
</tr>
</tbody>
</table>

Note: These estimates pertain to cluster randomized controlled trials and use the formulas in Tool 6.3 to estimate the minimum sample size requirement for the one-intervention group condition. The formulas underlying the tool are based on those reported in Bloom (2006). The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology. Available online at: http://www.mdrc.org/publications/437/full.pdf.

*For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).*
### Assumptions

**Standards of evidence**

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Significance level for null hypothesis test (generally 5%)</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>2. Two-tailed or one-tailed Test? (generally set at 2)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3. Statistical Power Standard (generally 80%)</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>4. MRI-ES (Minimum relevant impact in standard deviation units)</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Properties of the study sample**

<table>
<thead>
<tr>
<th>Property</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. PT (proportion of the sample allocated to the intervention condition)</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>6. R² (proportion of the variance in the outcome explained by background control variables)</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>7. Sample retention at follow-up (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

8. **Guestimated minimum required sample** (use Exhibit 6.2 to guage starting values)

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. Guestimated minimum required sample</td>
<td>394</td>
<td>176</td>
</tr>
</tbody>
</table>

9. **Estimated Minimum Required Sample Size**: If not equal to guestimated value in previous row, replace the guestimate with the value in this row.

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Estimated Minimum Required Sample Size</td>
<td>394</td>
<td>176</td>
</tr>
</tbody>
</table>

---

## Standards of Evidence

1. Significance level for null hypothesis test (two-tailed test) (generally 5%)  
   1. 0.05  
   2. 0.05  
   3. 0.05

2. Two-tailed or one-tailed Test? (generally set at 2)  
   1. 2  
   2. 2  
   3. 2

3. Statistical power standard (generally 80%)  
   1. 0.80  
   2. 0.80  
   3. 0.80

4. Standard deviation of the outcome measure (use empirical estimate whenever possible)  
   1. 4.00  
   2. 4.00  
   3. 4.00

## Properties of the study sample

5. $p_I$ (proportion of the sample allocated to the intervention condition)  
   1. 0.50  
   2. 0.50  
   3. 0.50

6. $R^2$ (proportion of the variance in the outcome explained by background control variables)  
   1. 0.00  
   2. 0.00  
   3. 0.00

7. Sample retention at follow-up (generally 80% or higher)  
   1. 100%  
   2. 100%  
   3. 100%

8. Number of control variables used in the analysis (most important for small samples)  
   1. 0  
   2. 0  
   3. 0

9. N (total sample size)  
   1. 394  
   2. 500  
   3. 1000

## Output

10. MDI (minimum detectable impact)  
    1. 1.132  
    2. 1.004  
    3. 0.709

11. MDI-ES (minimum detectable impact in standard deviation units)  
    1. 0.283  
    2. 0.251  
    3. 0.177
### Assumptions

#### Standards of evidence
1. Significance level for null hypothesis test (generally 5%)  
   - Benchmark: 0.05  
   - Alternative Scenario 1: 0.05  
   - Alternative Scenario 2: 0.05
2. Two-tailed or one-tailed Test? (generally set at 2)  
   - Benchmark: 2  
   - Alternative Scenario 1: 2  
   - Alternative Scenario 2: 2
3. Statistical Power Standard (generally 80%)  
   - Benchmark: 0.80  
   - Alternative Scenario 1: 0.80  
   - Alternative Scenario 2: 0.80
4. MRI-ES (Minimum relevant impact in standard deviation units)  
   - Benchmark: 0.2  
   - Alternative Scenario 1: 0.2  
   - Alternative Scenario 2: 0.2

#### Properties of the study sample
5. Nc (average number of analysis units per cluster)  
   - Benchmark: 20  
   - Alternative Scenario 1: 20  
   - Alternative Scenario 2: 20
6. Intraclass correlation (ICC or the proportion of the variance in the outcome that is between clusters)  
   - Benchmark: 0.10  
   - Alternative Scenario 1: 0.10  
   - Alternative Scenario 2: 0.00
7. Pr (proportion of the sample allocated to the intervention condition)  
   - Benchmark: 0.50  
   - Alternative Scenario 1: 0.50  
   - Alternative Scenario 2: 0.50
8. R² (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)  
   - Benchmark: 0.50  
   - Alternative Scenario 1: 0.00  
   - Alternative Scenario 2: 0.50
9. R² (proportion of the variance in the outcome explained by background control variables used in the cluster-level models)  
   - Benchmark: 0.20  
   - Alternative Scenario 1: 0.00  
   - Alternative Scenario 2: 0.00
10. Number of group level background control variables used  
    - Benchmark: 5  
    - Alternative Scenario 1: 0  
    - Alternative Scenario 2: 0
11. Sample retention for the analysis sample at follow-up (generally 80% or higher)  
    - Benchmark: 100%  
    - Alternative Scenario 1: 100%  
    - Alternative Scenario 2: 100%
12. Sample retention rate for the cluster level observations (generally 80% or higher)  
    - Benchmark: 100%  
    - Alternative Scenario 1: 100%  
    - Alternative Scenario 2: 100%

#### Estimated minimum required sample
(Use Table 6.2 to gauge starting values)

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Alternative Scenario 1</th>
<th>Alternative Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Guestimated minimum required sample</td>
<td>83</td>
<td>115</td>
<td>20</td>
</tr>
<tr>
<td>14. Estimated Minimum Required Number of Clusters</td>
<td>83</td>
<td>115</td>
<td>20</td>
</tr>
<tr>
<td>14. Estimated Minimum Required Number of Analysis Units</td>
<td>1651</td>
<td>2292</td>
<td>395</td>
</tr>
</tbody>
</table>


^ For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).
### Standards of evidence

1. Significance level for null hypothesis test (two-tailed test) (generally 5%)  
   - Version 1: 0.05  
   - Version 2: 0.05  
   - Version 3: 0.05

2. Two-tailed or one-tailed Test? (generally set at 2)  
   - Version 1: 2  
   - Version 2: 2  
   - Version 3: 2

3. Statistical power standard (generally 80%)  
   - Version 1: 0.80  
   - Version 2: 0.80  
   - Version 3: 0.80

### Properties of the study sample

1. N (Number of clusters randomized)  
   - Version 1: 35  
   - Version 2: 35  
   - Version 3: 35

2. n (Average number of analysis units per cluster)  
   - Version 1: 30  
   - Version 2: 30  
   - Version 3: 30

3. Intra-class correlation (ICC)  
   - Version 1: 0.10  
   - Version 2: 0.10  
   - Version 3: 0.10

4. $P_I$ (proportion of the sample allocated to the intervention condition)  
   - Version 1: 0.50  
   - Version 2: 0.50  
   - Version 3: 0.50

5. $R^2_I$ (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)  
   - Version 1: 0.50  
   - Version 2: 0.50  
   - Version 3: 0.00

6. $R^2_C$ (proportion of the variance in the outcome explained by background control variables used in the cluster-level models)  
   - Version 1: 0.20  
   - Version 2: 0.20  
   - Version 3: 0.20

7. Number of group level background control variables used (most important for small samples)  
   - Version 1: 1  
   - Version 2: 1  
   - Version 3: 1

8. Sample retention at follow-up (generally 80% or higher)  
   - Version 1: 100%  
   - Version 2: 50%  
   - Version 3: 100%

9. Sample Retention Rate for Students (generally 80% or higher)  
   - Version 1: 100%  
   - Version 2: 50%  
   - Version 3: 100%

10. Standard deviation of the outcome measure (use empirical estimate whenever possible)  
    - Version 1: 4.000  
    - Version 2: 4.000  
    - Version 3: 4.000

<table>
<thead>
<tr>
<th></th>
<th>Version 1</th>
<th>Version 2</th>
<th>Version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDI-ES</td>
<td>0.301</td>
<td>0.426</td>
<td>0.324</td>
</tr>
<tr>
<td>MDI- Natural units (MDI-ES*SD_{Outcome})</td>
<td>1.204</td>
<td>1.703</td>
<td>1.296</td>
</tr>
</tbody>
</table>

**Note:** The MDI-ES is computed using the formula in Bloom 2007, adjusted to incorporate sample retention parameters. This formula is presented in Appendix 6.1.
### Exhibit 6.1: Assumptions for Simple Random Assignment Studies Estimate the Effects of Mandatory Summer School for Low Achieving High School Students

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Persistence</th>
<th>Academic Achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standards of Evidence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance level of the null hypothesis test (probability of concluding there is a nonzero impact when there is not)</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Power (probability of obtaining a statistically significant impact estimate when the true impact is at least as great as the minimum relevant impact (MRI))</td>
<td>80%</td>
<td>80%</td>
</tr>
<tr>
<td>Minimum Relevant Impact (MRI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum Relevant Impact in natural units (MRI\textsubscript{Natural units})</td>
<td>5.00%</td>
<td>1.00</td>
</tr>
<tr>
<td>MRI expressed in standard deviations (MRI-ES)</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Properties of the Study Sample</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of the sample assigned to treatment condition (N\textsubscript{t})</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Mean of the outcome measure for &quot;business as usual condition&quot;</td>
<td>90%</td>
<td>9.00</td>
</tr>
<tr>
<td>Standard Deviation (Sdv) of the outcome measure for the business as usual condition</td>
<td>30%</td>
<td>4.00</td>
</tr>
<tr>
<td>Number of background control variables to be included in the analysis (N\textsubscript{v})</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Estimated proportion of the variance in the outcome measure explained by the background control variables (R\textsuperscript{2})</td>
<td>20%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Adjusted Minimum Relevant Impact (AMRI)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Minimum Relevant Impact in natural units (AMRI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= MRI ((1- R^2)^{.5})</td>
<td>5.59%</td>
<td>1.41</td>
</tr>
<tr>
<td>AMRI expressed in standard deviations (AMRI-ES)</td>
<td>0.19</td>
<td>0.35</td>
</tr>
<tr>
<td><strong>Minimum Required Sample Size to Meet Evidence Standards</strong></td>
<td>910</td>
<td>253</td>
</tr>
</tbody>
</table>

Note: The minimum required sample size is based on the formula in Tool 6.1.
### Exhibit 6.2: Illustrative Sample Size Requirements to Meet Commonly Accepted Standards of Evidence Requirements (80 Percent Power and 5 Percent Significance Level for the Null Hypothesis Test)

<table>
<thead>
<tr>
<th>Minimum relevant impact in standard deviation units (MRI-ES or AMRI-ES)</th>
<th>Proportion of the study sample assigned to the treatment group (p,)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>0.1</td>
<td>3158</td>
</tr>
<tr>
<td>0.2</td>
<td>790</td>
</tr>
<tr>
<td>0.3</td>
<td>351</td>
</tr>
<tr>
<td>0.4</td>
<td>197</td>
</tr>
<tr>
<td>0.5</td>
<td>126</td>
</tr>
<tr>
<td>0.6</td>
<td>88</td>
</tr>
<tr>
<td>0.7</td>
<td>64</td>
</tr>
<tr>
<td>0.8</td>
<td>49</td>
</tr>
<tr>
<td>0.9</td>
<td>39</td>
</tr>
<tr>
<td>1.0</td>
<td>32</td>
</tr>
</tbody>
</table>

#### Penalty for Unbalanced Design

| (N_{unbalanced}/N_{balanced}) | 1.00 | 1.04 | 1.19 | 1.56 | 2.78 |

Note: The minimum sample size is that required to be able to detect true impacts as small as that designated as the minimum relevant impact with 80 percent power and 95 percent confidence in situations where the study sample is randomly assigned to treatment or control condition. The formula for computing the minimum sample size required is the following:

\[
N = N_t + N_c = (2.81/\text{MRI-ES})^2/((N_t/N)*(1-N_t)/N))
\]

where

- N = total sample size for analysis
- \(N_t\) and \(N_c\) = the number assigned to the treatment and control groups, respectively
- 2.81 represents a multiplier that takes account of the statistical significance standard for the null hypothesis test (set at .05, two-tailed test) and the power level (80%) in this example.
- MRI-ES = the minimum relevant impact in standard deviation units the study needs to be able to detect at the designated standards of evidence (statistical power and precision)
Exhibit 6.3: Illustrative Minimum Number of Clusters Randomized to Achieve MDE-ES of .2 (.05 significance; 80% power; and 20 units per cluster), by variance explained by covariates and ICC

<table>
<thead>
<tr>
<th>Percent of Variance Explained</th>
<th>0.00</th>
<th>0.10</th>
<th>0.20</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only individual-level covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.3</td>
<td>265</td>
<td>262</td>
<td>259</td>
<td>256</td>
<td>254</td>
<td>251</td>
<td>248</td>
</tr>
<tr>
<td>ICC = 0.2</td>
<td>190</td>
<td>187</td>
<td>183</td>
<td>180</td>
<td>177</td>
<td>174</td>
<td>171</td>
</tr>
<tr>
<td>ICC = 0.1</td>
<td>116</td>
<td>111</td>
<td>107</td>
<td>104</td>
<td>100</td>
<td>97</td>
<td>93</td>
</tr>
<tr>
<td>ICC = 0.0</td>
<td>40</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>24</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Only cluster-level covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.3</td>
<td>265</td>
<td>241</td>
<td>217</td>
<td>194</td>
<td>170</td>
<td>146</td>
<td>123</td>
</tr>
<tr>
<td>ICC = 0.2</td>
<td>190</td>
<td>174</td>
<td>158</td>
<td>142</td>
<td>126</td>
<td>111</td>
<td>95</td>
</tr>
<tr>
<td>ICC = 0.1</td>
<td>115</td>
<td>107</td>
<td>99</td>
<td>91</td>
<td>83</td>
<td>75</td>
<td>67</td>
</tr>
<tr>
<td>ICC = 0.0</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Individual and cluster-level covariates, similar predictive power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC = 0.3</td>
<td>265</td>
<td>238</td>
<td>212</td>
<td>185</td>
<td>159</td>
<td>132</td>
<td>106</td>
</tr>
<tr>
<td>ICC = 0.2</td>
<td>190</td>
<td>171</td>
<td>152</td>
<td>133</td>
<td>114</td>
<td>95</td>
<td>76</td>
</tr>
<tr>
<td>ICC = 0.1</td>
<td>115</td>
<td>103</td>
<td>92</td>
<td>80</td>
<td>69</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td>ICC = 0.0</td>
<td>40</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>24</td>
<td>20</td>
<td>16</td>
</tr>
</tbody>
</table>

Note 1: These estimates are based on the formulas used in Tool 6.2. They pertain to cases where the units of assignment are the same as the units of analysis (e.g., analysis of student outcomes when students were randomly assigned to intervention or control condition or analysis of teacher outcomes when teachers were randomized.

Note 2: Other things equal, the number of sample clusters needed to achieve a larger (smaller) MDE-ES will can be approximated by multiplying the above numbers by the inverse of the (.2/MDE-ES*), where MDE-ES is the alternative minimum detectable effect size. For example, an MDE-ES of .1 would require samples approximately 4 times as large as those reported in this table, while an MDE-ES of .3 would require a sample .44 times as large as those reported in this table.
### Assumptions

#### Standards of evidence
1. Significance level for null hypothesis test (generally 5%)  
   - Benchmark: 0.05  
   - Alternative Scenario 1: 0.05  
   - Alternative Scenario 2: 0.05
2. Two-tailed or one-tailed Test? (generally set at 2)  
   - Benchmark: 2  
   - Alternative Scenario 1: 2  
   - Alternative Scenario 2: 2
3. Statistical Power Standard (generally 80%)  
   - Benchmark: 0.80  
   - Alternative Scenario 1: 0.80  
   - Alternative Scenario 2: 0.80
4. MRI-ES (Minimum relevant impact in standard deviation units)  
   - Benchmark: 0.2  
   - Alternative Scenario 1: 0.2  
   - Alternative Scenario 2: 0.2

#### Properties of the study sample
5. Nc (average number of analysis units per cluster)  
   - Benchmark: 20  
   - Alternative Scenario 1: 20  
   - Alternative Scenario 2: 20
6. Intraclass correlation (ICC or the proportion of the variance in the outcome that is between clusters)  
   - Benchmark: 0.10  
   - Alternative Scenario 1: 0.10  
   - Alternative Scenario 2: 0.00
7. Pr (proportion of the sample allocated to the intervention condition)  
   - Benchmark: 0.50  
   - Alternative Scenario 1: 0.50  
   - Alternative Scenario 2: 0.50
8. R² (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)  
   - Benchmark: 0.50  
   - Alternative Scenario 1: 0.00  
   - Alternative Scenario 2: 0.50
9. R² (proportion of the variance in the outcome explained by background control variables used in the cluster-level models)  
   - Benchmark: 0.20  
   - Alternative Scenario 1: 0.00  
   - Alternative Scenario 2: 0.00
10. Number of group level background control variables used  
    - Benchmark: 5  
    - Alternative Scenario 1: 0  
    - Alternative Scenario 2: 0
11. Sample retention for the analysis sample at follow-up (generally 80% or higher)  
    - Benchmark: 100%  
    - Alternative Scenario 1: 100%  
    - Alternative Scenario 2: 100%
12. Sample retention rate for the cluster level observations (generally 80% or higher)  
    - Benchmark: 100%  
    - Alternative Scenario 1: 100%  
    - Alternative Scenario 2: 100%

#### Estimated Minimum Required Numbers of Clusters
14. Estimated Minimum Required Number of Clusters: If not equal to guestimated value in previous row, replace the guestimate with the value in this row.  
    - Benchmark: 83  
    - Alternative Scenario 1: 115  
    - Alternative Scenario 2: 20

#### Estimated Minimum Required Number of Analysis Units
14. Estimated Minimum Required Number of Analysis Units: If not equal to guestimated value in previous row, replace the guestimate with the value in this row.  
    - Benchmark: 1651  
    - Alternative Scenario 1: 2292  
    - Alternative Scenario 2: 395


*For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).*
### Study and Outcome Measure

<table>
<thead>
<tr>
<th>Sample Case</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reading test (grade 1)</td>
<td>Teacher Retention</td>
<td>Reading test scores</td>
<td>Math test scores</td>
<td>Reading/Treatment</td>
<td>Reading/Pooled treatments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. ( N_{\text{CLUSTERS}} )</td>
<td>35</td>
<td>158</td>
<td>960</td>
<td>252</td>
<td>240</td>
<td>38</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>2. ( P_T )</td>
<td>0.50</td>
<td>0.56</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.45</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>3. ( N_{\text{IC}} )</td>
<td>30</td>
<td>17</td>
<td>1</td>
<td>21</td>
<td>20</td>
<td>71</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>4. ICC</td>
<td>0.10</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>5. ( R_{\text{c}}^2 )</td>
<td>0.50</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>6. ( R_{\text{c}}^2 )</td>
<td>0.20</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>7. ( N_{\text{CVar}} )</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>8. SDV-Outcome</td>
<td>1.000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### Minimum Detectable Impacts

| | 10. MDI-ES | 0.301 | 0.159 | 0.162 | 0.119 | 0.124 | 0.283 | 0.212 | #NUM! | #NUM! | #NUM! |
| | MDI-Natural Units | 0.301 | 0.159 | 0.162 | 0.119 | 0.124 | 0.283 | 0.212 | #NUM! | #NUM! | #NUM! |

Note: The MDI-ES calculations in this worksheet are based on the formula used in Tools 6.2 and 6.4, depending whether or not the outcome pertains to the units that were randomly assigned or to subunits within the assignment clusters.

Information on the various studies can be found in the following documents:

1. Dynarski et al. (2007)....
2. Glazerman et al. 2006...teacher induction
3. PD Math TWG 2007
4. AIR 2004... PD math NOTE: 3-group design.
5. James-Burdamy et al. 2006. Reading comp per intervention
6. James-Burdamy et al. 2006. Reading comp per intervention/ pooled across 4 interventions

Made up data for now. This table will be completed using data from the study reports.
### Properties of the study sample

<table>
<thead>
<tr>
<th></th>
<th>One Intervention Group</th>
<th>Two Intervention Groups</th>
<th>Three Intervention Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Number of treatment conditions</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2.</td>
<td>Nc (average number of analysis units per cluster)</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3.</td>
<td>Intraclass correlation (ICC or the proportion of the variance in the outcome that is between clusters)</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>4.</td>
<td>P&lt;sub&gt;T&lt;/sub&gt; (proportion of the sample allocated to the intervention condition)</td>
<td>0.50</td>
<td>0.33</td>
</tr>
<tr>
<td>5.</td>
<td>R&lt;sup&gt;2&lt;/sup&gt; (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>6.</td>
<td>Number of group level background control variables used</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>7.</td>
<td>Sample retention for the analysis sample at follow-up (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>8.</td>
<td>Sample retention rate for the cluster level observations (generally 80% or higher)</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Estimated Minimum Sample Sizes

<table>
<thead>
<tr>
<th></th>
<th>One Intervention Group</th>
<th>Two Intervention Groups</th>
<th>Three Intervention Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.</td>
<td>Clusters for Each Treatment/Control Comparison</td>
<td>83</td>
<td>91</td>
</tr>
<tr>
<td>14.</td>
<td>Clusters Total Across All Treatment and Control Groups</td>
<td>83</td>
<td>122</td>
</tr>
<tr>
<td>15.</td>
<td>Analysis Units for Each Treatment Control Comparison</td>
<td>1651</td>
<td>1823</td>
</tr>
<tr>
<td>16.</td>
<td>Analysis Units for All Treatment-Control Comparisons</td>
<td>1651</td>
<td>2430</td>
</tr>
</tbody>
</table>

Note: These estimates pertain to cluster randomized controlled trials and use the formulas in Tool 6.3 to estimate the minimum sample size requirement for the one-intervention group condition. The formulas underlying the tool are based on those reported in Bloom (2006). The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology. Available online at: http://www.mdrc.org/publications/437/full.pdf.

<sup>a</sup>For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).
## Assumptions

### Standards of evidence

1. Significance level for null hypothesis test (generally 5%)  
   - Version 1: 0.05  
   - Version 2: 0.05
2. Two-tailed or one-tailed Test? (generally set at 2)  
   - Version 1: 2  
   - Version 2: 2
3. Statistical Power Standard (generally 80%)  
   - Version 1: 0.80  
   - Version 2: 0.80
4. MRI-ES (Minimum relevant impact in standard deviation units)  
   - Version 1: 0.2  
   - Version 2: 0.3

### Properties of the study sample

5. $P_I$ (proportion of the sample allocated to the intervention condition)  
   - Version 1: 0.50  
   - Version 2: 0.50
6. $R^2$ (proportion of the variance in the outcome explained by background control variables)  
   - Version 1: 0.50  
   - Version 2: 0.50
7. Sample retention at follow-up (generally 80% or higher)  
   - Version 1: 100%  
   - Version 2: 100%

8. **Guestimated minimum required sample** (use Exhibit 6.2 to gauge starting values)  
   - Version 1: 394  
   - Version 2: 176

9. **Estimated Minimum Required Sample Size**: If not equal to guestimated value in previous row, replace the guestimate with the value in this row.  
   - Version 1: 394  
   - Version 2: 176

---

### Tool 6.2: Minimum Detectable Effect Calculator for Two-group, One-level Samples

<table>
<thead>
<tr>
<th>Standards of evidence</th>
<th>User Input and Program Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Significance level for null hypothesis test (two-tailed test) (generally 5%)</td>
<td>0.05 0.05 0.05</td>
</tr>
<tr>
<td>2. Two-tailed or one-tailed Test? (generally set at 2)</td>
<td>2 2 2</td>
</tr>
<tr>
<td>3. Statistical power standard (generally 80%)</td>
<td>0.80 0.80 0.80</td>
</tr>
<tr>
<td>4. Standard deviation of the outcome measure (use empirical estimate whenever possible)</td>
<td>4.00 4.00 4.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Properties of the study sample</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5. $p_t$ (proportion of the sample allocated to the intervention condition)</td>
<td>0.50 0.50 0.50</td>
</tr>
<tr>
<td>6. $R^2$ (proportion of the variance in the outcome explained by background control variables)</td>
<td>0.00 0.00 0.00</td>
</tr>
<tr>
<td>7. Sample retention at follow-up (generally 80% or higher)</td>
<td>100% 100% 100%</td>
</tr>
<tr>
<td>8. Number of control variables used in the analysis (most important for small samples)</td>
<td>0 0 0</td>
</tr>
<tr>
<td>9. N (total sample size)</td>
<td>394 500 1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10. MDI (minimum detectable impact)</td>
<td>1.132 1.004 0.709</td>
</tr>
<tr>
<td>11. MDI-ES (minimum detectable impact in standard deviation units)</td>
<td>0.283 0.251 0.177</td>
</tr>
</tbody>
</table>
### Assumptions

#### Standards of evidence
1. Significance level for null hypothesis test (generally 5%)
   - Benchmark: 0.05
   - Alternative Scenario 1: 0.05
   - Alternative Scenario 2: 0.05
2. Two-tailed or one-tailed Test? (generally set at 2)
   - Benchmark: 2
   - Alternative Scenario 1: 2
   - Alternative Scenario 2: 2
3. Statistical Power Standard (generally 80%)
   - Benchmark: 0.80
   - Alternative Scenario 1: 0.80
   - Alternative Scenario 2: 0.80
4. MRI-ES (Minimum relevant impact in standard deviation units)
   - Benchmark: 0.2
   - Alternative Scenario 1: 0.2
   - Alternative Scenario 2: 0.2

#### Properties of the study sample
5. Nc (average number of analysis units per cluster)
   - Benchmark: 20
   - Alternative Scenario 1: 20
   - Alternative Scenario 2: 20
6. Intraclass correlation (ICC or the proportion of the variance in the outcome that is between clusters)
   - Benchmark: 0.10
   - Alternative Scenario 1: 0.10
   - Alternative Scenario 2: 0.00
7. Pr (proportion of the sample allocated to the intervention condition)
   - Benchmark: 0.50
   - Alternative Scenario 1: 0.50
   - Alternative Scenario 2: 0.50
8. R² (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)
   - Benchmark: 0.50
   - Alternative Scenario 1: 0.00
   - Alternative Scenario 2: 0.50
9. R² (proportion of the variance in the outcome explained by background control variables used in the cluster-level models)
   - Benchmark: 0.20
   - Alternative Scenario 1: 0.00
   - Alternative Scenario 2: 0.00
10. Number of group level background control variables used
    - Benchmark: 5
    - Alternative Scenario 1: 0
    - Alternative Scenario 2: 0
11. Sample retention for the analysis sample at follow-up (generally 80% or higher)
    - Benchmark: 100%
    - Alternative Scenario 1: 100%
    - Alternative Scenario 2: 100%
12. Sample retention rate for the cluster level observations (generally 80% or higher)
    - Benchmark: 100%
    - Alternative Scenario 1: 100%
    - Alternative Scenario 2: 100%

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Alternative Scenario 1</th>
<th>Alternative Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Guestimated minimum required sample</td>
<td>83</td>
<td>115</td>
<td>20</td>
</tr>
<tr>
<td>14. Estimated Minimum Required Number of Clusters</td>
<td>83</td>
<td>115</td>
<td>20</td>
</tr>
<tr>
<td>14. Estimated Minimum Required Number of Analysis Units</td>
<td>1651</td>
<td>2292</td>
<td>395</td>
</tr>
</tbody>
</table>


*For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).*
### Standards of evidence

1. Significance level for null hypothesis test (two-tailed test) (generally 5%)
   - Version 1: 0.05
   - Version 2: 0.05
   - Version 3: 0.05
2. Two-tailed or one-tailed Test? (generally set at 2)
   - Version 1: 2
   - Version 2: 2
   - Version 3: 2
3. Statistical power standard (generally 80%)
   - Version 1: 0.80
   - Version 2: 0.80
   - Version 3: 0.80

### Properties of the study sample

1. N (Number of clusters randomized)
   - Version 1: 35
   - Version 2: 35
   - Version 3: 35
2. n (Average number of analysis units per cluster)
   - Version 1: 30
   - Version 2: 30
   - Version 3: 30
3. Intra-class correlation (ICC)
   - Version 1: 0.10
   - Version 2: 0.10
   - Version 3: 0.10
4. P_I (proportion of the sample allocated to the intervention condition)
   - Version 1: 0.50
   - Version 2: 0.50
   - Version 3: 0.50
5. R^2 (proportion of the variance in the outcome explained by background control variables included in the individual sample member analysis)
   - Version 1: 0.50
   - Version 2: 0.50
   - Version 3: 0.00
6. R_c^2 (proportion of the variance in the outcome explained by background control variables used in the cluster-level models)
   - Version 1: 0.20
   - Version 2: 0.20
   - Version 3: 0.20
7. Number of group level background control variables used (most important for small samples)
   - Version 1: 1
   - Version 2: 1
   - Version 3: 1
8. Sample retention at follow-up (generally 80% or higher)
   - Version 1: 100%
   - Version 2: 50%
   - Version 3: 100%
9. Sample Retention Rate for Students (generally 80% or higher)
   - Version 1: 100%
   - Version 2: 50%
   - Version 3: 100%
10. Standard deviation of the outcome measure (use empirical estimate whenever possible)
    - Version 1: 4.000
    - Version 2: 4.000
    - Version 3: 4.000

<table>
<thead>
<tr>
<th>MDI-ES</th>
<th>Version 1</th>
<th>Version 2</th>
<th>Version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.301</td>
<td>0.426</td>
<td>0.324</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MDI- Natural units (MDI-ES*SdOutcome)</th>
<th>Version 1</th>
<th>Version 2</th>
<th>Version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.204</td>
<td>1.703</td>
<td>1.296</td>
<td></td>
</tr>
</tbody>
</table>

Note: The MDI-ES is computed using the formula in Bloom 2007, adjusted to incorporate sample retention parameters. This formula is presented in Appendix 6.1.