Program Evaluation Toolkit
Module 4, Chapter 3: Sampling Plan

Regional Educational Laboratory Central

From the National Center for Education Evaluation at IES

Speaker 1:

Welcome to the third chapter of Module 4. In this chapter, you will explore methods for determining sample size and then create a sampling plan for your evaluation.

There are two types of data you might collect: quantitative data and qualitative data. Quantitative data are numeric and answer, "how much" and "to what extent" questions. Qualitative data are non-numeric and answer, "why" and "how" questions. Module five will cover these two types of data in more depth. Before you begin collecting either type of data, you should determine the adequate sample size you will need to feel confident that your results are not due merely to chance. If you collect data from too few individuals, you might have skewed findings. If you collect data from too many individuals, you might waste resources by gathering data you do not need. Qualitative data collection is less dependent on large sample sizes and tends to involve nonrandom sampling types. Quantitative data collection requires larger sample sizes and tends to involve random sampling types.

Qualitative data collection tends to involve nonrandom sampling. Data are often collected through qualitative methods until saturation is achieved. *Saturation* is the point at which data collected begin to yield no new information. For example, if the AMMP! evaluation team interviews teachers about student barriers to completing math homework, the team may reach a point at which the teachers identify no new barriers during the interviews. The team should then conclude that it has reached saturation and stop collecting data.

This chapter focuses on quantitative data collection in which you use either a random sampling type or purposive sampling. In both cases, you should determine the sample size needed to detect an effect or group difference related to your program. In the case of consecutive, convenience, or snowball sampling, you do not need to be as concerned with sample size because you want to obtain the largest sample possible, given your resource constraints.

To identify the sample size for random sample or purposive data collection, you should first consider the unit of measurement. The unit of measurement is the level at which you plan to collect data. For example, if you are examining whether a program had an effect on student achievement, and you plan to look at the test scores of individual students, then your unit of measurement is students.

In some cases, you may need to protect individual identities by looking at a phenomenon through a larger lens. For instance, even if a measure of student achievement is of primary importance to your evaluation, you may need to obtain average measures of classroom or school achievement rather than test scores of individual students to protect students' identities. Consider whether

your unit of measurement can be at the individual level without personally identifying students, or whether it should be at a higher level such as a classroom, school, or district.

After you identify the unit of measurement, the next step is to identify the sampling unit. The sampling unit is the level at which you identify individuals for the sample. Sometimes this is the same as the unit of measurement. For instance, if you want to collect data from teachers, then teachers are both the unit of measurement and the sampling unit. However, the sampling unit does not have to be the same as the unit of measurement. In fact, it will often be useful if they are not the same. In clustered random sampling, for example, the sampling unit is not the same as the unit of measurement. When you want to obtain measurement from many individuals (for example, teachers, students, and parents), it is often useful for the sampling unit to be at a higher, group level (for example, schools are sampled to identify teachers, students, and parents from the same school).

As you plan an evaluation in which you will use random sampling, you can use the *Sample Size Workbook*, a Microsoft Excel tool available on the resources page of the website, to calculate the sample size. However, you should not use this tool for stratified or clustered random sampling because determining an appropriate sample size in these cases is much more complicated. More advanced tools are available that can handle stratification and clustering. For instance, the PowerUp! tool is available for free download, although this tool is best used in consultation with an expert in education statistics.

The *Sample Size Workbook* can still be useful when you are interested in measurements of many individuals within a school. For instance, suppose your evaluation will measure principals, teachers, and students in a context in which each school has one principal. Because there are many more teachers and students than principals at each school, if you plan your evaluation to have enough schools to answer the evaluation questions related to principals, you will usually also have enough teachers and students to answer the evaluation questions related to teachers and students.

But how do you determine how many units to sample to be confident that your results will be similar to what you might see if you collected data from the entire population? You need to consider two procedures that tell you how large a sample you need to be confident in your results. Procedures for determining the appropriate sample size for a quantitative evaluation are based on inferential statistics. Inferential statistics allow you to draw conclusions about a population from a sample. Module 7 covers inferential statistics in more detail. The two main procedures that make up inferential statistics are *confidence intervals* and *hypothesis tests*.

Confidence intervals are calculated for statistics generated from a sample of a larger population, such as a sample mean. These intervals depict a range of values that would contain the true value for the population a certain proportion of times if you repeated your sampling over and over again., When that proportion needs to be high—that is, when you need to have a higher level of confidence that your sample-based estimate is typical of the whole population—the range of values is typically wider and a larger sample size is typically required.

Standard practice is to calculate 95 percent confidence intervals. A 95 percent confidence interval tells you that, were you to repeat your sampling infinitely many times, each time calculating a new average and confidence interval around that average, 95 percent of the intervals you created would contain the population's true average.

In the AMMP! example, when determining the sample size needed to answer the evaluation question "To what extent do students complete homework with better than 80 percent accuracy?" the evaluation team wants a sample large enough to have 95 percent confidence that there is only a small difference between the population and sample proportions that meet this criterion. Guidance on how to use the *Sample Size Workbook* to determine the necessary sample size appears later in this chapter.

The first step in *hypothesis testing* involves writing a null hypothesis of no effect, which is a statement that asserts that there will be no difference between a control and treatment group involved in an evaluation. The purpose of hypothesis testing is to determine whether the information in the sample is sufficient to reject the null hypothesis of no effect. In the AMMP! example, the evaluation team may want to know if, on average, students participating in AMMP! perform differently on an end-of-year math assessment from students not participating in AMMP!. This would address the evaluation question "How do AMMP! participants' scores on high school math placement tests compare to nonparticipants' scores?" The null hypothesis would be that, on average, students participating in AMMP! and those not participating in AMMP! perform exactly the same. The team might seek to determine if the data provide enough evidence to reject this hypothesis.

All conclusions in inferential statistics are subject to the possibility that they are in error. Hypothesis testing is most often concerned with the error of rejecting the null hypothesis when it is true—that is, asserting there is an effect when there truly is not one. Typically, hypothesis testing procedures limit the probability of making this error to 5 percent.

If AMMP! is effective at increasing students' scores on the math placement test, the evaluation team wants to be able to reject the null hypothesis. *Statistical power* is the probability of rejecting the null hypothesis when a particular alternative hypothesis is true. For instance, the alternative hypothesis might be that students participating in AMMP! score 2 points higher on the placement exam, on average, than students not participating in AMMP!.

Statistical power will increase if the sample size is larger. Standard practice is to choose the sample size to ensure that statistical power is at least 80 percent.

Although you can use either confidence intervals or statistical power for any type of evaluation, you should use confidence intervals for descriptive and correlational designs and statistical power for quasi-experimental designs (QEDs) and randomized controlled trials (RCTs). See module 3 for more details about these evaluation designs. Confidence intervals may be more appropriate when there is no obvious comparison group or hypothesis. For example, district staff might simply want to examine the number of tutors trained on AMMP!. However, statistical power is most useful when data analysis involves comparing outcomes for two groups, as occurs in most QEDs and RCTs. For example, the school district using AMMP! might want to compare

the math achievement of their students to those students in a similar school district that is not using AMMP!.

The next step in planning the sample size is to identify the type of quantitative data you will collect to inform your evaluation questions. Although there are many types of quantitative data, this chapter focuses on the two most common types: continuous and binary.

Continuous data can take on a wide range of possible values. Examples include student test scores, years of teaching experience, and schoolwide percentage of students eligible for the National School Lunch Program.

On the other hand, binary data can take on only two values. Examples include an exam with a pass or fail score, course completion, graduation, or college acceptance.

So how might you use continuous and binary data? Often, with continuous data, you want to calculate an average or mean. For instance, the AMMP! evaluation team might calculate the average math test score for students who did and did not participate in the program. With binary data, you want to calculate a percentage. For example, the AMMP! evaluation team might be interested in the percentage of students in the middle school who passed a math exam after participating in the program.

When planning the sample size for an evaluation with continuous data, the required sample size will usually depend on how spread out the data are. Standard deviation is a commonly used measure of spread. (Module 7 includes more detail on standard deviation.) A standard deviation tells you how spread out your data are within a given sample. When there are two groups of interest, your sample size calculations assume that the standard deviation is the same in the two groups.

Over the next few slides on sample size planning, when the data are continuous, we will discuss distance in terms of standard deviation units. Doing so will simplify sample size planning. We will also use the *Sample Size Workbook* to calculate sample sizes for different scenarios.

It is worth noting that when there are two groups of interest, calculations will be based on the assumption that the standard deviation in the two groups is the same.

To begin, let's focus on estimating a single mean for a continuous variable. For this scenario, let's suppose the AMMP! evaluation team wants to answer the evaluation question "To what extent are parents satisfied with their students' homework completion?"

To answer this question, the evaluation team decides to ask AMMP! parents to rate their satisfaction with their students' math homework completion on a scale of 0 percent to 100 percent satisfied. Suppose that previous surveys indicated that the standard deviation of parents' satisfaction ratings tended to be around 4 percentage points. The team could determine this standard deviation by reading previously published literature about parental satisfaction surveys.

The evaluation team must decide how close the average rating of the sample of parents needs to be to the true average rating for all parents of AMMP! students in order for the evaluation to be

useful. One way the team can decide is to read previously published literature to learn what level of difference in satisfaction is considered practically meaningful or policy relevant. Another way the team can decide is to informally interview a few principals. Suppose the team learns that principals regard differences in parent satisfaction scores of less than 1 percentage point as trivial but regard differences of more than 1 percentage point as worthy of attention. Then it would be reasonable for the team to decide that the difference between the sample mean and the population mean should not exceed 1 percentage point. This means that the desired interval width is 2 percentage points. Because sample size planning measures distance in standard deviation units, the team divides the desired width of 2 percentage points by the standard deviation of 4 percentage points to determine that the necessary interval width is 0.5 standard deviations. To meet this standard deviation interval, the team wants to estimate the true average rating for all parents in the district to within plus or minus 1 percentage point.

To do this calculation in the *Sample Size Workbook*, choose the "Confidence interval for a single mean" heading, under the "Confidence Intervals" tab. When you enter the desired interval width of 0.5 in the appropriate place in the spreadsheet, you obtain the required sample size of 62, the number of parents the AMMP! evaluation team will need to address the evaluation question.

Now let's suppose the AMMP! evaluation team wants to answer the evaluation question "To what extent are parents of students participating in AMMP! more or less satisfied with their students' homework completion, compared to parents of nonparticipating students?"

To address this question, the evaluation team needs to compare the average response of parents with students participating in AMMP! to the average response of parents with students not participating in AMMP! The team wants to estimate the difference between the true average rating for all AMMP! parents and the true average rating for all non-AMMP! parents to within plus or minus 2 percentage points.

Again, the team assumes that the standard deviation of parent ratings is 4 percentage points.

In this scenario, the evaluation team needs to specify what proportion of the sample will be in each group. It is usually preferable to sample an equal number from each group, if possible. However, in this case, because there are fewer students in AMMP!, the team samples 25 percent of AMMP! parents and 75 percent of non-AMMP! parents.

To do this calculation in the *Sample Size Workbook*, use the "Confidence Intervals" tab and look under the "Confidence interval for comparing two means" heading. Enter 0.5 for the desired interval width, and 0.75 for the proportion of the sample that will come from the first group, and you obtain the necessary sample size. The team needs to sample a total of 328 parents, 82 from families participating in AMMP! and 246 from the nonparticipating families.

Now let's consider an example with binary data. Suppose the AMMP! evaluation team still wants to answer the evaluation question "To what extent are parents satisfied with their students' homework completion?" However, only binary data are available, in which parents indicated "yes" or "no."

The evaluation team wants to estimate the proportion of "yes" responses, and they must decide how close the sample estimate of this proportion needs to be to the true proportion for the evaluation to be useful. Suppose the team makes this decision by consulting principals, who tell the team it is important to make sure that the sample proportion and the true proportion do not differ by more than .04, or 4 percentage points. This means that the desired confidence interval width is 8.

The team wants to estimate the proportion of all parents who are satisfied to within plus or minus 4 percentage points, 8 in total.

With binary data, instead of assuming a standard deviation, the team needs a rough estimate of the proportion of parents who will answer "yes." In this case, the team assumes that this value is 60 percent.

The team finds that it needs to sample 577 parents to feel confident that the team will collect data from enough parents to get an accurate sample. Note that, with binary data, the sample size requirements will be larger.

To do this calculation in the *Sample Size Workbook*, use the "Confidence Intervals" tab and look under the "Confidence interval for a single proportion" heading and enter the desired interval width and the guessed at value of the proportion of "yes" answers. Once you do this, you find that the evaluation team needs 577 parents in the evaluation.

As a general rule, you need to have a larger sample to have precise estimates of proportion relative to evaluations with continuous data.

Now let's suppose that, to answer the evaluation question "To what extent are parents of students participating in AMMP! are more or less satisfied with their students' homework completion, compared to parents of nonparticipating students?" the AMMP! evaluation team wants to compare the proportion of AMMP! parents who indicate "yes" with the proportion of non-AMMP! parents who indicate "yes."

In this case, the team wants to estimate the difference between the true proportion of AMMP! parents who are satisfied and the true proportion of non-AMMP! parents who are satisfied.

The team assumes that the proportion of satisfied AMMP! parents is about 70 percent, that the proportion of satisfied non-AMMP! parents is about 50 percent, and that 25 percent of the sample will be AMMP! parents.

The team wants to estimate the difference between the proportions to within plus or minus 8 percentage points, which means they need a confidence interval width of 16. As discussed previously, there are many ways the evaluation team might make this determination. In this case, the team holds informal discussions with principals, who consider 8 percentage points to be the smallest practically meaningful difference (in other words, differences of less than 8 percentage points would not be worthy of attention, according to most principals). The team finds that it needs to sample 673 parents.

To do this calculation in the *Sample Size Workbook*, navigate to the section of the spreadsheet with the heading "Confidence interval for comparing two proportions." Enter 16 for the desired interval width, 0.7 for the estimate of the proportion of *yes* responses from AMMP! parents, 0.5 for the proportion of *yes* responses from non-AMMP! parents, and 0.25 for the proportion of the sample that will be AMMP! parents. You can see that the required sample size is 705 parents.

This is a large number. As noted earlier, it is best to plan for equally sized groups when possible. The team now changes the problem and assumes that 50 percent of the sample will be AMMP! parents. In this case, the required sample size is only 553.

Now let's suppose that the AMMP! evaluation team wants to conduct a quasi-experimental design (or QED) or randomized controlled trial (or RCT) that compares two groups after making a statistical adjustment to account for preexisting differences between the groups. This decision implies that the team wants to make a causal comparison that applies to the whole district and that the team will use power analysis for sample size planning, unlike previous scenarios. The sample size planning approach discussed in this scenario is most appropriate for planning an RCT. However, the results are also approximately correct for planning a QED that adjusts for preexisting differences between groups, provided that those differences are not overly large.

Suppose that the evaluation team wants to answer the evaluation question "How do AMMP! participants' scores on high school math placement tests compare to nonparticipants' scores?" To make this comparison, the team must first identify the smallest difference between AMMP! students' scores and non-AMMP! students' scores that would be sufficient to justify keeping the program. The specification should be in standard deviation units and is known as the *minimum relevant effect size* (MRES). Effect size is a measurement of the strength of the association between an intervention and outcome. In the AMMP! context, effect size is defined as the difference between the average scores of AMMP! students and non-AMMP! students, divided by the standard deviation of those scores. The larger the effect size, the stronger the relationship. In some cases, the effect size might indicate a small but important relationship, so the team wants to determine the smallest relationship that the study should be able to detect, using MRES. When the outcome is a measure of student achievement (such as a test score), many evaluations in education research assume a MRES between 0.20 and 0.33.

The AMMP! evaluation team also needs to determine the proportion of the sample that will have participated in AMMP! in order to conduct the power analysis. The team will likely have this information on hand, which is the number of students who will have participated in AMMP! divided by the total number of students eligible to participate.

The AMMP! evaluation team wants to determine the causal impact of AMMP! participation on high school math placement test scores. Of course, different students will obtain different scores on the test. The extent of these differences is called the variation in the scores. If AMMP! is effective, it will explain some of the variation in these scores. However, other variables will also explain variation in the scores. These variables may include eligibility for the National School Lunch Program, each student's test scores in math from previous years, homework completion percentage, gender, and race/ethnicity.

The team can simply compare means to answer the evaluation question, but if additional variables that explain math score variation are measured and included in regression models (discussed in Module 7), it will be possible to conduct an equally effective evaluation with a smaller sample size. To determine the required sample size, the evaluation team needs to estimate the percentage of variation in math scores that can be explained by the additional variables included in regression models. Previous research has shown that preintervention scores on a test similar to the outcome measure (for example, a previous math test when the outcome is a math test score) can explain a large amount of variance, between 50 and 80 percent. Demographic covariates such as gender or eligibility for the National School Lunch Program can be expected to explain a smaller amount of variance, between 20 and 50 percent. Note that the variance-explained numbers cannot be added. In other words, if the evaluation measures both preintervention test scores and demographic covariates, these variables will likely still explain 50 to 80 percent of variance, but not 100 percent.

Going back to the AMMP! example, navigate to the "Power-Mean Difference" tab of the *Sample Size Workbook*. The evaluation team might assume a standardized MRES of 0.25, because the team will have access to pre-AMMP! test scores. The team assumes that 60 percent of the variance (proportion of variance explained by covariates [R^2]) in math scores can be explained by other variables. The team also assumes that 75 percent of the sample is not participating in AMMP!.

If you go to the *Sample Size Workbook* and enter the assumed values, you find that the evaluation team needs 268 students in their evaluation.

Now that you have learned sampling considerations and used the *Sample Size Workbook*, let's apply them to your own evaluation questions. Choose one or more of your questions and use the *Sampling Plan for Evaluation Questions* handout, available on the resources page of the website, to draft a sampling plan. Use the *Sample Size Workbook* to complete the sampling plan. Creating a sampling plan may take you up to 45 minutes. PowerUp! is a more advanced tool that can help you choose a sample size in more complicated cases, such as for stratified or clustered random sampling.

After you draft a sampling plan, you are ready to move to the next step of the program evaluation cycle: Module 5. In that module, you will review components of data quality and learn to identify existing data sources. For users who wish to conduct more complicated sampling plans but take into account variation among schools or classrooms we recommend utilizing PowerUp!, which will provide users with the ability to also take these considerations into account.

This handout was prepared under Contract ED-IES-17-C-0005 by Regional Educational Laboratory Central, administered by Marzano Research. The content does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.