Facilitator notes: This presentation provides an overview of the different evaluation designs that can be used to conduct a program evaluation.

Facilitator notes: For this presentation, we have identified a number of learning objectives.

By the end of this presentation, you will be able to:

- Explain evaluation design
- Describe the differences between types of evaluation designs
- Identify the key elements of each type of evaluation design
- Understand the key considerations in selecting a design for conducting an evaluation of your AmeriCorps program

Facilitator notes: This presentation will cover the following:

- What is evaluation design?
- CNCS's evaluation continuum
- How to select an appropriate evaluation design for your program
- Key elements of each type of evaluation design
- Evaluation resources and tools

To facilitate your understanding of the information presented, we also provide a few exercises that we'll be doing at various points during the presentation.

Evaluation is the use of research methods to assess a program's design, implementation, and/or outcomes. Every evaluation is essentially a research or discovery project. Evaluation looks at the results of your investment of time, expertise, resources and energy, and compares those results with what you said you wanted to achieve in your program's logic model. Your research may be about determining how effective your program or intervention is overall, which parts of it are working well, and which components need adjustment.

For more information on logic models, CNCS grantees should refer to the module, "How to Develop a Program Logic Model" located on the Knowledge Network.

Before we discuss evaluation design, we first want to provide you with a brief overview of performance measurement and program evaluation and, in particular, how these activities differ from one another. While both performance measurement and program evaluation are considered measurement activities, the two activities serve different purposes.

Some of you may already know what performance measurement is because it's an activity that you should already be doing in your program. Performance measurement is the ongoing monitoring and reporting of program accomplishments and progress toward its pre-established goals. For many programs, this includes collecting data on the specific activities carried out and the direct products and services produced by your program's activities (outputs and outcomes). Performance measurement data help you understand what level of performance is achieved by the program/intervention.

Program evaluation, on the other hand, is an in-depth research activity that answers specific questions or tests hypotheses about program processes and/or program outcomes. The results enable you to arrive at a judgment of whether the intervention or a specific component of the intervention works or does not work as expected, and also what adjustments may be needed to improve the program. A key difference between performance measurement and program evaluation is that program evaluation helps you understand and explain why you're seeing the program results.

Logic models can be used as a tool in both performance measurement and evaluation. Logic models can help you plan your performance measurement activities by identifying which components of your program (resources, activities, outputs, outcomes) to include in your performance measurement. Logic models also can help you identify indicators and measures of progress or performance that align with program components. Logic models also can inform program evaluation by helping you make informed decisions about what to evaluate, when to evaluate, and how you will evaluate. Your logic model can be used as a tool to help you focus your evaluation with respect to the following:

- Identify questions you want or need answered about your program
- Identify which aspects of your program to evaluate (e.g., will you evaluate a subset or all of your AmeriCorps activities? Will you evaluate your program's short-term outcomes?)
- Determine the appropriate evaluation design (e.g., will you use a process or an impact evaluation design, or a combination of both?)
- Identify what information to collect
- Identify measures and data collection methods
- Determine an appropriate timeframe for your evaluation

It is important to keep in mind that performance measurement and program evaluation are not mutually exclusive. Grantees already engaged in performance measurement activities can build on that work as they plan for a program evaluation. For example, let's say your program is already collecting data to monitor and report on your program's progress toward achieving its expected outcomes for program beneficiaries. If your program decides to conduct an evaluation, it may be that you continue to collect the SAME outcomes data, but use it to answer specific questions about your program and perhaps collect the same data from a comparison group that does not participate in your program.

Grantees who want to learn more about program evaluation can refer to the Knowledge Network webpage.

https://www.nationalserviceresources.gov/evaluation-americorps

Grantees who want to learn more about performance measurement can refer to the Knowledge Network webpage.

https://www.nationalserviceresources.gov/npm/ac

This diagram illustrates CNCS's overall developmental approach. It shows that evidence falls along a continuum with the understanding that identifying an evidence-based program model requires organizational capacities that correspond to an organization's life cycle. The key building blocks for generating evidence are shown in the diagram. The first step is identifying a strong program design by gathering evidence that supports the intervention to be used. During this initial process, it is helpful to develop a logic model which clearly communicates the central model of your program. We will discuss logic models in more detail later in this presentation. It also is recommended that the program be piloted during this initial step to ensure its effective implementation prior to expanding the program more widely.

Once a strong program design has been identified, the second building block is ensuring the effective full implementation of the program. Efforts should be made to document program processes, ensure fidelity to the central program model, evaluate program quality and efficiency, and establish continuous process improvement protocols. Much of these activities can be supported through the identification and regular monitoring of performance measures.

The next level in the continuum is assessing the program's outcomes. This process involves developing indicators for measuring outcomes, possibly conducting one of the less rigorous outcome evaluation designs, such as a single group pre-post design to measure program outcomes, and conducting a thorough process evaluation. We will discuss what these types of evaluation designs entail later in this presentation.

One step further in the continuum is obtaining evidence of positive program outcomes by examining the linkages between program activities and outcomes. Programs at this level of the continuum will have performed multiple pre- and post-evaluations and conducted outcome evaluations using an independent evaluator.

Finally, the highest level of evidence allows a program to make the claim of being evidence-based by attaining strong evidence of positive program outcomes. At this level, programs have established the causal linkage between program activities and intended outcomes/impacts. Programs at this level have completed multiple independent evaluations using strong study designs, such as a quasi-experimental evaluation using a comparison group or an experimental, random assignment design study. Many of these programs also have measured the cost effectiveness of their program compared to other interventions addressing the same need.

Based on this understanding of a continuum of evidence, a strong program design, sound performance measures, and the identification of measureable program outcomes are a fundamental starting point for building evidence of effectiveness. Consequently, attempts to generate experimental evidence before earlier developmental work has been completed is not recommended and may result in wasting valuable resources. As an agency, CNCS continues to develop a funding strategy that will create a portfolio of programs reflecting a range of evidence levels (e.g., strong, moderate, preliminary) that are appropriate to the program's life cycle and investment of public dollars. CNCS sees value in infusing evaluative thinking and knowledge into every phase of a program's life cycle – program development, implementation, improvement, and replication/scaling.

Now that we have provided a clear understanding of CNCS's developmental approach to evaluation, we will turn our attention to discussing what is evaluation design. If the results of your evaluation are to be reliable, you have to give the evaluation a structure that will tell you what you want to know. That structure is the evaluation's design, and it includes why the evaluation is being conducted, what will be measured, who will participate in the evaluation, when and how data will be collected, what methods will be used to collect the data, and whether a comparison group is appropriate and how feasible it is to identify one. The evaluation design you choose depends on what kinds of questions your evaluation is meant to answer. We will talk more about how research questions determine the evaluation design later in the presentation.

The appropriate design for evaluating a program will largely depend upon certain considerations. Most important to selecting an appropriate evaluation design is a clear and detailed understanding of your program model. Also, it is important to have a clear understanding of what is the primary purpose or goal of the evaluation. For example, do you want to focus on the process of program implementation (what your program does and how well you do it) or on the outcomes achieved (what difference did your program make), or both? Also, the specific evaluation questions you want the evaluation to answer will help determine which type of evaluation design you should choose. Another important consideration is the resources available for the evaluation, such as staff time, outside expertise, and funding. Finally, the evaluation requirements laid out by your funder, such as CNCS, are also a key consideration in which design you select. We are going to discuss each of these key considerations in more detail next.

The first consideration is selecting an evaluation design that aligns with your program model. As you begin to plan for an evaluation of a program or intervention, it is essential that there be a clear and comprehensive mapping of the program or intervention itself. Thus, a useful first step in planning an evaluation should be to clarify and confirm your program's operations or processes and intended outcomes by developing a logic model. If your program has already developed a logic model, then you might only need to review the existing model and possibly update or refine it to reflect your current program operations and goals.

Let's talk about what a program logic model is:

A program logic model is a detailed visual representation of your program and its theory of change that communicates how your program works, the resources you have to operate your program, the activities you carry out, and the outcomes you hope to achieve. Your program logic model should clearly communicate how your program works by depicting the intended relationships among program components. Key program components consist of:

- Inputs or resources - which are considered essential for a program's activities to occur. They may include any combination of human, financial, organizational, and community-based resources that are available to a program and used to carry out a program's activities.
- Activities – which are the specific actions that make up your program or intervention. They reflect processes, tools, events, and other actions that are used to bring about your program's desired changes or results.
- Outputs – what a program's specific activities will create or produce, providing evidence of service delivery (e.g., the number of beneficiaries served or the number of children improving reading scores).
- Outcomes - the specific changes that may result from a program's activities or intervention. A program's outcomes fall along a continuum, ranging from short- to long-term results (e.g., an increase in knowledge of healthy food choices, a decrease in delinquency rates, or an increase in literacy).

Logic models are typically read from left to right, employing an if-then sequence among key components. A generic example is shown here. It reads, if your program has these inputs or resources, then it can carry out these activities. If your program carries out these activities, then it can produce these outputs. If your program has produced these outputs, then it will achieve these outcomes.

In addition, we can think of a logic model as essentially having two "sides." The **process** side focuses on a program's implementation or its planned work – inputs/resources, activities, and outputs (direct products). The **outcomes** side of the logic model describes the expected sequence of changes that the program is to accomplish, which can be short-term, medium-term, and/or long-term changes. The outcomes side reflects what difference the program intends to make.

On this slide, we present an example logic model for a hypothetical literacy program. Logic models come in many different forms and for this example, we use CNCS's NOFO template for a program logic model.

This hypothetical program is designed to address the low literacy rates of elementary school students in California. In this example program, a school district in California is implementing a literacy program in several of their elementary schools. The program involves AmeriCorps members delivering one-on-one tutoring using research-based interventions for elementary school students who are not reading at pre-established targets. Once students are scoring on benchmark, they are graduated from the tutoring program.

The logic model we present here is a visual summary of this program. We'll read through the logic model together, starting from the left column and moving progressively to the right.

On the left side, we begin with the program's investments, referred to as **inputs** or resources. For this literacy program, this includes

- Funding
- Program staff
- AmeriCorps members
- Non-AmeriCorps volunteers
- Research for identifying evidence-based literacy interventions to be used

Moving to the next column, if this program has these inputs, then it can carry out its planned **activity** which in this case is:

One-on-one tutoring to students scoring below expected benchmarks on assessment tests

This activity will create or produce the following **output**, which refers to the product or evidence that the activity was carried out:

- # students receiving tutoring assistance

The next column refers to the **short-term outcomes** of the program. These are the immediate changes that are expected to result from program services and activities. In the short-term, this hypothetical program expects to see an increase in the number of students scoring at or above benchmark on literacy assessments, as well as improved student self-efficacy.

Moving to the **medium-term outcomes** column, these outcomes reflect changes in behavior or action that that are expected to occur after short-term outcomes have been achieved. The medium-term outcome for this hypothetical program is an increase in the number of students reading on grade-level.

The last column refers to **long-term outcomes**. If students are able to read on grade-level, the expected long-term outcome for this program is that students will be able to maintain grade-level proficiency in reading.

We will use this example logic model throughout our presentation.

For a more detailed explanation of logic models, CNCS grantees should refer to the module, “How to Develop a Program Logic Model” located on the Knowledge Network.

Once your program has a clear understanding of its program’s operations, processes, and intended outcomes through the development of a logic model, a second consideration is the type of evaluation you want to complete on your program. Just as your program needs to have a specific purpose and scope, so does your evaluation. Thus, an important consideration in determining your evaluation design is defining the purpose and scope of your evaluation. Each evaluation should have a *primary* purpose around which it can be designed and planned, although it may have several other purposes. The stated purpose of the evaluation drives the expectations and sets the boundaries for what the evaluation can and cannot deliver.

In defining the purpose of the study, it is helpful to identify why the evaluation is being done, what you want to learn from the evaluation findings, and how the information collected and reported by the study will actually be used and by whom. For example, are program staff trying to understand how to operate the program more efficiently or identify barriers or constraints to implementation? Or does your program need to produce evidence that it is meeting its intended outcomes? Will the results be used by program staff to make changes to the program’s implementation? Could the results be used to generate interest from other funders? In general, defining a specific purpose for your evaluation will allow you to set parameters around the design you use, data you collect and methods you will use.

Questions about why your evaluation is being done and how the information will be used should be discussed among a variety of program staff, and any other individuals who may be involved in the evaluation to ensure there is consensus as to what the evaluation will accomplish.

A third consideration in selecting an evaluation design is which research questions you want the evaluation to address. Turning back now to a logic model, on this slide we present an example of how your logic model can be used to help you focus your evaluation by narrowing in on the primary question or questions you want to address.

The graphic above provides an example of the types of questions that may be asked of each component in a logic model:

- Questions related to **inputs** ask, “Are resources adequate to implement the program?”
- Questions related to **activities** ask, “Are activities delivered as intended?”
- Questions related to **outputs** ask, “How many, how much was produced?”
- Questions related to **outcomes** ask**,** “What changes occurred as a result of the program?”

As you can see at the bottom of this graphic, in order to answer each of these questions, indicators (i.e., the evidence or information that represents the phenomenon in question) and their data sources will need to be identified. When developing research questions based on your program’s logic model, ensure that questions are stated in a way that is clear and measurable in order for them to be answerable.

Once you identify the questions you want to answer, this information will guide your selection of the type of evaluation design-- process or outcomes-- required to answer your questions.

A fourth consideration in selecting an evaluation design is what resources (staff time, funding, evaluation expertise) are available to carry out the evaluation. Because most programs have limited resources that can be put towards an evaluation, it is important to note that it is not necessary to evaluate every aspect of your program all at once as depicted in your logic model. Your evaluation can have a narrow focus (e.g., only address questions about one of your program's service activities and desired outcomes) or it can have a broader focus (e.g., address questions about each of your program's service activities and desired outcomes), depending on the information you hope to gain from your evaluation and the resources you have available.

It is important to note that evaluation is not a one-time activity. Program evaluation should be thought of as part of a series of activities over time that align with the life cycle of your program. Ultimately a series of evaluations will build upon one another and generate more knowledge and evidence of your program's effectiveness over time. *Facilitator may want to refer back to slide 6 to reiterate that CNCS sees value in infusing evaluative thinking and knowledge into every phase of a program's life cycle – program development, implementation, improvement, and replication/scaling.*

We noted also that your funders' evaluation requirements are another consideration in selecting an evaluation design. We will discuss CNCS' evaluation design requirements for large and small grantees towards the end of this presentation.

Now that we've presented the key considerations in selecting an evaluation design, we will discuss some of the different types of evaluation designs in more detail.

First, please recall that a logic model essentially has two "sides," a process and an outcomes side. Similar to what is reflected in the logic model, a process evaluation focuses on answering questions about your program's inputs, activities, and outputs, while an outcome evaluation answers questions about what changes occurred as a result of your program, as measured by your short, medium, and long-term outcomes. Process or implementation evaluations answer questions such as "What did you do and how well did you do it?" while outcome evaluations answer questions such as "What difference did your program make?" Next, we are going to discuss each of these designs starting with the process evaluation.

A process evaluation can be used to document what a program is doing and to what extent and how consistently the program demonstrates fidelity to the program's logic model. The results of a process evaluation are most often used to change or improve the program.

Process evaluations are able to address research questions about why a project is or is not successful, which can be very helpful for program staff and stakeholders because the results are useful for improving program practices. To answer the types of research questions associated with a process evaluation generally a comparison group is not necessary. The collection of both qualitative and quantitative data through interviews, surveys, and program administrative data is usually preferred.

Additionally, process evaluations mostly rely on simple descriptive statistics (means, frequencies, etc.) and do not require advanced statistical methods.

It is also worth noting that the results of process evaluations are usually not generalizable, meaning that they cannot be applied to similar program models being implemented in locations other than those participating in the evaluation.

Process evaluations can be used to answer one or more of a number of questions about a program, including, but not limited to:

- Is the program being implemented as designed or planned?
- Is the program being implemented the same way at each site?
- Is the program reaching the intended target population with the appropriate services at the planned rate and "dosage"?
- Are there any components of the program that are not working well? Why or why not?
- Are program beneficiaries generally satisfied with the program? Why or why not?
- Are the resources adequate for the successful implementation of the program?

The program will explore these research questions by examining the following types of data: Program and school level administrative data and site visits to the schools to examine the fidelity of program implementation. The site visits will consistent of observations of literacy intervention with individual students, interviews with school staff and administration, and several focus groups with teachers and students.

For the analysis, information gathered through these various sources will be compared across sources to identify themes that may emerge based on the consistency of responses. Information gathered through interviews and focus groups can sometimes be confirmed through the use of other quantitative data sources, such as administrative records. For example, a teacher may comment in a focus group that their school struggled with implementing the program because so many of their families regularly move during the school year due to employment issues, so students don’t end up receiving the full intervention. Other teachers and the school principal also point out this same barrier to implementation. Administrative records showing large numbers of students transferring into and out of the school during the school year can be used to confirm their point.

To show how a process evaluation might be designed, we offer a facilitated group exercise using the literacy program presented earlier.

As explained previously, a school district in California is implementing a literacy program using an existing model. The program wants to know, “Is the literacy program being implemented consistent with the program’s logic model and theory of change?” We will develop together a basic example of a general approach to designing a process evaluation to answer this question.

First, we should turn back to the logic model of the program (Handout #1). We want to think about the program and what are our major design considerations:

- What to measure
- Who to include in the evaluation
- When and how often data will be collected

- What methods will be used to collect data

For this exercise, we provide you with a crosswalk that can be created to help you think through the process of determining the major design considerations for the given research question. We encourage all of you to participate and provide your input as we design a process evaluation for this program together.

The first column begins the research question under study. The main research question for the process evaluation concerns whether the literacy program is being implemented as designed.

Moving to the next column, what might be some potential indicators for assessing fidelity to the program model?

Below are possible responses:

- member use of program curriculum during tutoring sessions
- the duration of sessions
- student attendance rates at sessions

Next, who or from what sources might we be able to obtain this information?

Possible responses include:

- AmeriCorps members
- Evaluator

Depending on the audience's responses on the indicators, other possibilities include teachers, parents, school administrators, etc.

Moving to the next column, when and who would collect this information?

Possible responses include:

- Member tutoring logs quarterly
- Quarterly observation by the evaluator using structured observation protocols

Once the data have been gathered, what approaches would you use to analyze the data?

Possible responses include:

- Simple descriptive statistics can be generated from the quantitative data such as frequencies on the use of the curriculum and averages on the duration of workshop and participant attendance rates.
- Qualitative data that have been collected may be thematically coded and analyzed.

Taken together, analyses of all the collected data are then used to assess the extent to which the program was implemented as designed.

This slide is only intended to be an example for the facilitator who may or may not elect to present this to the audience. The group exercise will likely yield a different set of indicators, data sources, timing of data collection, and methods for data analysis than the examples listed here.

For this next exercise, we are going to divide up into small groups and develop a process evaluation for a literacy program. We will be using the same literacy program, but focus on a different research question. Again we provide you with a crosswalk to help you think through the process of determining the major design considerations for the given research question. In considering your approach to this research question on the satisfaction of program beneficiaries, it is important to consider who are the beneficiaries of the literacy program (Students, parents, teachers, schools). These different groups should be kept in mind when filling out the crosswalk. Similar to our earlier exercise, we'd like for you to fill out this crosswalk for a process evaluation of the literacy program. Once everyone has completed the exercise, we will share the various answers that the groups came up with.

(Asking the whole group now) What are some possible ways to assess the satisfaction levels of each group of program beneficiaries?

The first column begins the research question under study. The main research question for the process evaluation concerns whether program beneficiaries are satisfied with the literacy program.

Moving to the next column, what might be some potential indicators for assessing program satisfaction by beneficiary group (students, parents, teachers, schools)?

Below are possible responses:

- Satisfaction levels
- Continued student attendance

Next, who or from what sources might we be able to obtain this information?

Possible responses include:

- Parents
- Teachers
- School administrators
- AmeriCorps members

Moving to the next column, when and who would collect this information?

Possible responses include:

- Parent survey sent home with students by the schools at the end of tutoring
- Focus groups with teachers at the end of each semester facilitated by the evaluator
- Interviews with school administrators at the end of the school year conducted by the evaluator
- Member tutoring logs quarterly

Once the data have been gathered, what approaches would you use to analyze the data?

Possible responses include:

- Simple descriptive statistics can be generated from the quantitative survey data such as average satisfaction levels and participant attendance rates.
- Qualitative data that have been collected may be thematically coded and analyzed.

Taken together, analyses of all the collected data are then used to assess beneficiary satisfaction with the program.

This slide is only intended to be an example for the facilitator who may or may not elect to present this to the audience. The group exercise will likely yield a different set of indicators, data sources, timing of data collection, and methods for data analysis than the examples listed here.

An outcome or impact evaluation can be used to determine the results or effects of a program. These types of evaluations generally measure changes in program beneficiaries' knowledge, attitudes, or behaviors thought to result from the program. Outcome or impact evaluations are best implemented after a program has had sufficient time to mature and is no longer undergoing refinement of its central model.

More rigorous outcome evaluations, such as quasi-experimental and experimental design studies, include a comparison group against which to measure changes in program beneficiaries. These types of outcome evaluation designs are referred to as impact evaluations. The use of a comparison group provides additional evidence that observed changes in program beneficiaries were due to the program or intervention. Thus, impact evaluations are better able to measure or estimate the **impact** of the program on beneficiaries. These types of studies typically require quantitative data collection and often employ advanced statistical methods for analyzing data.

As previously explained, the more rigorous outcome evaluations, such as quasi-experimental and experimental design studies, include a comparison or control group, which is a group of individuals that either receive a different intervention than the one being evaluated or no intervention at all. A comparison or control group is necessary for deriving an estimate of the program’s impact by comparing the amount of change or improvement between comparison/control groups and those who participated in the program. The term comparison group is associated with a quasi-experimental design and the term control group is used when the evaluation employs an experimental design.

It is important to note that a comparison group is not just individuals who do not participate in the program. It’s important to be thoughtful in your selection of a comparison group to ensure that they are as similar as possible to those individuals enrolled in your program. This is key to ensuring that evaluation findings are unbiased, valid, and reliable. Including a comparison group enables you to answer specific questions related to causality – such as, what would have happened to people if they did not receive the intervention your program offers (i.e., whether the observed changes can be attributed to your intervention). In most cases, you will want an experienced evaluator to use statistical matching to ensure that your comparison group is as similar as possible to the group of program beneficiaries in your evaluation.

We will discuss how comparison and control groups are different later in the presentation.

Outcome evaluations can be used to answer one or more of a number of questions about a program, including, but not limited to:

- Are there differences in outcomes for program beneficiaries compared to those not in the program?
- Did all types of program beneficiaries benefit from the program or only specific subgroups?
- Did the program change beneficiaries’ knowledge, attitude, behavior, or condition?

As mentioned earlier, under the umbrella of outcome evaluation are a number of more specific designs. The main differences between quasi-experimental and experimental design studies, which are required for large grantees, and non-experimental design studies, which are acceptable designs for small grantees, is the use of a comparison group. Because of their use of a comparison group, only quasi-experimental and experimental evaluations can be considered impact evaluations because they provide more reliable evidence that changes in outcomes are due to the program itself. Next, we will discuss each of these types of designs.

We are going to begin our discussion with a presentation on several less rigorous outcome evaluation designs called non-experimental evaluations. We later will discuss quasi-experimental and experimental designs.

Although these approaches are available and may be suitable for a program in its initial stages of evaluation activity, CNCS does not recognize these designs as fulfilling the outcome evaluation requirement for large grantees. Some examples of these types of designs include the single group post design, which examines program beneficiaries after they receive program services, and the single group pre-post design, which examines program beneficiaries both before and after they receive program services. In these designs, no comparison group is used against which to measure change over time. The retrospective study design is another less rigorous evaluation design, in which previous program beneficiaries are asked to provide their opinion on the effects of the program services they received.

A vehicle which is commonly employed by AmeriCorps programs is member surveys, in which AmeriCorps members are asked questions on their service experiences and satisfaction levels. Members may also be asked to provide an opinion on the results of the program services they helped deliver. While member surveys may be useful for making decisions about program improvement and can even be a valuable source of data for use in a process evaluation, member surveys are not considered to be an outcome evaluation design and are not suitable for estimating program impacts on service beneficiaries.

While these designs are generally easier to complete, are low cost, and often do not require additional evaluation expertise outside of the program, they tend to produce less reliable results because they offer lots of opportunities to introduce bias (i.e., other influences) into study findings. Most importantly, these types of designs cannot produce results which can be considered attributable to the program. For example, an education program may find that students participating in a tutoring program experienced increases in their literacy skills based on measures taken at the beginning and end of the school year. However, most children should experience normal increases in their literacy skills over time whether an intervention is provided or not because learning is a naturally, cumulative process. Only when children in the program are compared to either an existing education benchmark or a comparison/control group will we be able to estimate whether the program actually produced greater gains in literacy than what would normally have occurred without the intervention.

A pre-post evaluation design is different than conducting a pre-post test for performance measurement. A pre-post test for performance measurement would generally compare outputs or performance measures before and after the intervention, whereas an evaluation would be more focused on measuring changes in outcomes further along the causal chain, such as changes in behavior.

Next we are going to discuss the two evaluation designs required to be completed by large grantees, quasi-experimental and experimental design.

It is not uncommon for people to confuse quasi-experimental and experimental designs because both types of designs include a comparison (or non-program) group against which to measure the outcomes of program beneficiaries. Quasi-experimental designs employ various methods to identify a similar group to the program participants. Experimental designs or Randomized Controlled Trials employ random assignment techniques to determine which program applicants will receive program services and which will receive some alternative intervention or no intervention. The control group in an experimental design may receive no services or alternative or delayed services. The use of random assignment in experimental designs creates as equal groups as possible by ensuring that there are no differences on average between the program and control groups. The process of randomly assigning program applicants for experimental design studies is complex and generally requires specific tailoring to each program's unique application and intake process. For this reason, random assignment is best conducted using a professional evaluator with experience completing these types of evaluations.

For quasi-experimental evaluation designs, statistical matching and/or propensity scoring is recommended to ensure that the program and comparison groups are as similar as possible when outcomes are compared. In the matching process, one or more comparison group members is identified and matched with a program beneficiary based on similar observable characteristics. This matching process enables a comparison of outcomes among program beneficiaries and comparison group members who are similar to one another. This often requires collecting a range of demographic data prior to or at the time of program entry in order to complete the matching process. Also, the proper matching of cases can be very complex, so it is best conducted by a professional evaluator with experience completing these types of evaluations.

There are several important differences between quasi-experimental and experimental evaluations. While quasi-experimental designs do not require the random assignment of program applicants, an important disadvantages of these types of evaluations is that it can be quite challenging for programs to identify a comparison group that is reasonably similar to program beneficiaries. Importantly, because the program and comparison groups are different and there are multiple approaches available for attempting to equate the study groups in the analysis process, the results of these types of studies are considered less rigorous and therefore can be more questionable. An important advantage of quasi-experimental evaluations, however, is that they can be less labor intensive and expensive because there is less need for intensive monitoring as with experimental evaluations.

There also are several benefits and challenges to experimental evaluations. Experimental designs are the most rigorous option available, so the findings from these studies are generally highly regarded and relied upon. There are several drawbacks, however, to experimental designs. They often require that a program increase its recruitment efforts to be able to have enough qualified applicants to form a control group. Because applicant acceptance is randomly determined, program staff may be dissatisfied with their lack of input into the program selection process. These types of designs can be more labor intensive and expensive because of the need for regular monitoring of study beneficiaries to ensure that applicants who are assigned to the program group actually participant in the program as intended and that applicants who are assigned to the control group do not receive program services or services from another similar program.

If someone asks, “Isn’t random assignment unethnical?” Answer: Generally, random assignment is an ethical and fair way to accept applicants into a program. When a program has more applicants than can be served at any one time, random assignment is actually a very fair way to ensure that everyone has an equal chance of being accepted to the program. Many evaluations are able to address concerns about the random assignment process by: 1) staggering treatment to ensure that all eligible applicants ultimately participate in the program; 2) using blocked randomization to ensure the required mix of program participants are accepted to the program (by age, grade, gender, etc.); and/or 3) allowing a small number of exceptions to the randomization process due to the special circumstances of a few individual applicants.

In specific instances, it may be impractical or unethical to use random assignment to identify program participants. For example, if individuals are court mandated to serve in a program, it would be unethical for the program not to serve every eligible applicant. In other cases, programs may be required to serve certain types of applicants that should not participate in the random assignment process. For example, many schools are required to serve special needs students or lose critical funding. Therefore, these students should not participate in the random assignment process. However, when a program is overprescribed (i.e., has more applicants than can be served at one time) random assignment or a lottery system is one of the fairest ways to determine which applicants are accepted to the program.

To provide further clarity on quasi-experimental and experimental studies, we are going to conduct a group exercise in which we will ask you to divide up into small groups and develop an outcomes design for the literacy program. We will designate some groups to develop a quasi-experimental design study and others to develop an experimental design study. Both types of designs should focus on answering the research question, “What impact does the literacy intervention program have on student reading levels relative to a comparison group of students?”

As with designing the process evaluation for this program, we want to think about our major design considerations:

- What to measure
- Who to include in the evaluation
- When and how often data will be collected
- What methods will be used to collect data

For this exercise, we will now work together to determining the major design considerations for this research question. We encourage everyone to participate and provide your input as we design an outcome evaluation for this program together.

The first column lists the research question under study. The main research question for the outcome evaluation asks what impact the literacy program has on students’ reading levels relative to a comparison group.

Moving to the next column, what might be some potential indicators for assessing fidelity to the program model?

Below is a possible response:

- student literacy assessment tests

Next, who or from what sources might we be able to obtain this information?

Possible responses include:

- Students enrolled in the program and students enrolled in a similar school that does not deliver the literacy program
- Within the same program, students receiving one-on-one tutoring and students receiving a different small group interventions

The main difference between a quasi-experimental design study and an experimental design study is the type of comparison group that is identified for evaluation, with the experimental design using a control group and the QED using a statistically matched comparison group.

Once the intervention and comparison groups have been identified, we ask when and who would collect this information? Most likely, the evaluator will collect the data at two time points (pre- and post-intervention). What are possible time points for collecting these data?

Possible responses includes:

- At the beginning of the semester before the program begins and at the end of the semester after students in the one-on-one tutoring group have received several months of tutoring
- At the beginning of the school year before the program begins and at the end of the school year

Finally, after the data have been gathered, how will the data be analyzed?

Below is a possible response:

- Statistical tests (in this case, difference-in-differences methods) can be used to compare program students with their matched comparison group by subtracting the average outcome (gain) in the comparison group from the average outcome (gain) in the intervention group.

Such analyses may show that, on average, students participating in the one-on-one tutoring program have higher rates of increase in reading improvement and reach benchmark reading levels at higher rates than students at schools without a similar literacy intervention.

This slide is only intended to be an example for the facilitator who may or may not elect to present this to the audience. The group exercise will likely yield a different set of outcomes of interest and their measurement, data sources, timing of data collection, and methods for data analysis than the examples listed here

For this next exercise, we are going to divide up into small groups and develop an outcome evaluation for the same literacy program, but focus on a different research question. Please use the handout of this same slide as you work together in your group. For this exercise, the main research question for the outcome evaluation asks what impact the literacy program has on students' self-efficacy relative to a comparison group. Self-efficacy refers to students' beliefs in their ability to succeed. We predict that students who receive reading assistance from the tutoring program will experience a greater increase in self-efficacy than students who do not participate in the program. The outcome of interest is student self-efficacy.

Once everyone has completed the exercise, we will share the various answers that the groups came up with.

(Asking the whole group now) What might be some potential indicators for assessing student self-efficacy?

Below is a possible response:

- self-efficacy questionnaire for children (there are existing, evidence-based instruments for measuring children's self-efficacy)

Moving to the next column, who or from what sources might we be able to obtain this information?

Possible responses include:

- Students enrolled in the program and students enrolled in a similar school that does not deliver the literacy program

The main difference between a quasi-experimental design study and an experimental design study is the type of comparison group that is identified for evaluation, with the experimental design using a control group and the QED using a statistically matched comparison group.

Once the intervention and comparison groups have been identified, the evaluator will collect the data at two time points (pre- and post-intervention). What are possible time points for collecting these data?

Possible responses includes:

- At the beginning of the semester before the program begins and at the end of the semester after students in the one-on-one tutoring group have received several months of tutoring
- At the beginning of the school year before the program begins and at the end of the school year

Finally, after the data have been gathered, how will the data be analyzed?

Below is a possible response:

- Statistical tests (in this case, difference-in-differences methods) can be used to compare program students with their matched comparison group by subtracting the average outcome (gain) in the comparison group from the average outcome (gain) in the intervention group.

Such analyses may show that, on average, students participating in the one-on-one tutoring program have higher rates of increase in self-efficacy than students at schools without a similar literacy intervention.

This slide is only intended to be an example for the facilitator who may or may not elect to present this to the audience. The group exercise will likely yield a different set of outcomes of interest and their measurement, data sources, timing of data collection, and methods for data analysis than the examples listed here.

Now that we have presented on the different evaluation designs, we want to conclude this presentation by discussing the evaluation designs as they pertain to CNCS' evaluation requirements. It is important to note that CNCS has different evaluation requirements for large and small recompeting grantees in terms of which evaluation design they may use to assess their programs. Large grantees are those receiving annual CNCS funds of $500,000 or more. Small grantees are those receiving annual CNCS funds of less than $500,000. You should note which type of design is required for your program.

Here we provide a list of resources on evaluation design that you may find helpful.

Does anyone have any questions?