NRS Myth Busters
Evaluation Research Guide

By:

Larry Condelli
Stephanie Cronen
Natalia Pane
Dahlia Shaewitz

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street, NW
Washington, DC 20007

This guide was prepared for the project:

Enhancing and Strengthening Accountability in Adult Education
Contract # ED-VAE-10-O-0107

For:

U.S. Department of Education
Office of Vocational and Adult Education
Division of Adult Education and Literacy

Cheryl Keenan, Director
Division of Adult Education and Literacy

Jay LeMaster, Program Specialist
Division of Adult Education and Literacy

June 2013

Contents

List of Exhibits

Introduction

Adult educators have been plagued by misinformation and negative perceptions about their students and programs. Myths such as these are common: *Adult education students never stay long enough to learn. NRS assessments do no inform teachers and instruction. Adult education students are just dropouts who have failed.* Due to the implementation of the National Reporting System (NRS) in 2000, states and adult education programs now have the means to counter these myths, describe teacher, instruction and services, student characteristics and program and student achievements. The NRS requirements to collect data on student participation and outcomes created large data sets that allowed states and local programs to describe who adult education students are, their attendance and participation, and the outcomes they achieve. The data created a valuable source of information for studying the relationships among program components, teachers, and student outcomes. These data can also help states and programs manage more effectively and help inform the field about what works to improve programs and promote student learning.

The Myth Busters' Evaluation Research Guide provides a detailed model and approach to assist state and local adult education programs to use their NRS data to "bust" these myths and to conduct research related to adult education students, teachers, or programs. The guide presents basic concepts of evaluation research, including developing research questions, research design, data analysis and reporting. This guide is the twelfth in a series of guides designed to assist states with implementing National Reporting System (NRS) requirements, improve data quality, and use NRS data to promote program improvement and supports national training conducted in June and July 2013.

The NRS support project staff at the American Institutes for Research developed all the NRS guides through OVAE funded projects that support the NRS. Readers interested in further information about the NRS and more information on data quality and the use of NRS data for program management and improvement should consult NRSWeb, the project website, at http://www.nrsweb.org/pubs/#trainingGuides, to obtain copies of these resources. The website also has training materials for all previous guides.

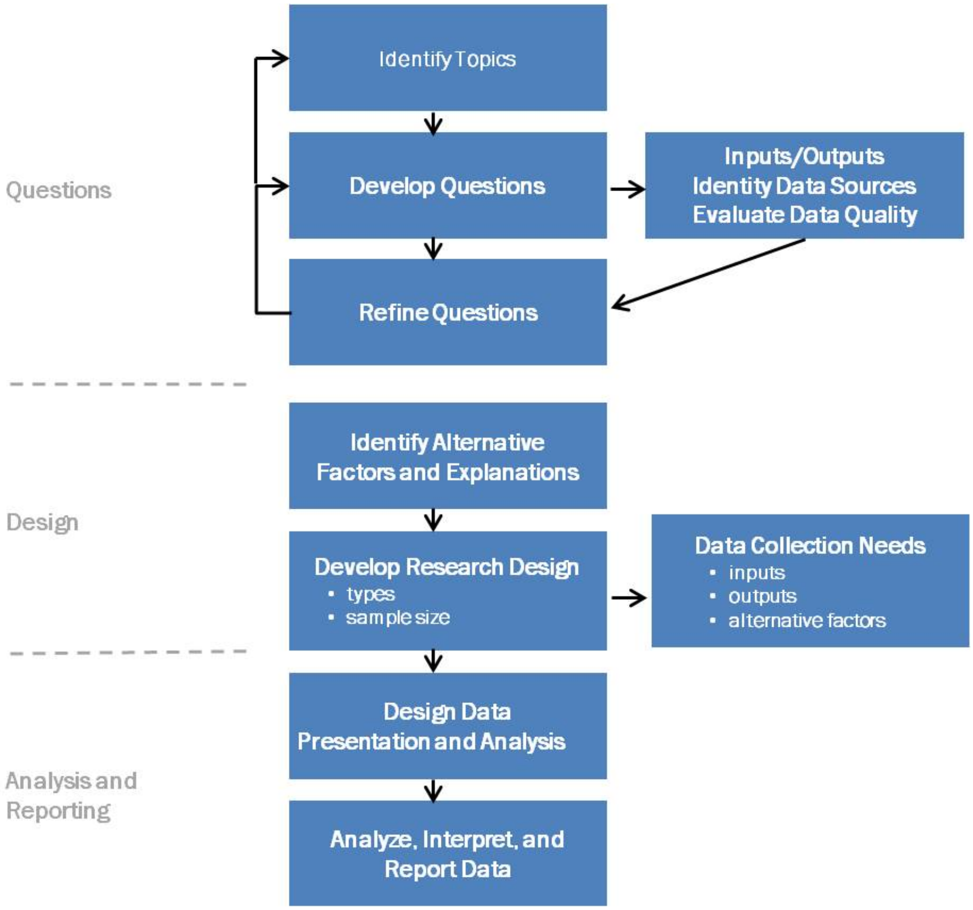
Evaluation Research Planning Model

The Myth Buster research planning model breaks the evaluation process into seven steps. Exhibit 1 shows the model, which is divided into three parts: developing research questions, research design, and analysis and reporting. The research **questions** process entails identifying issues or topics, developing questions to address the issues, and refining the questions. Identifying alternative factors and controls needed is also part of this process. The research **design** process includes determining the type of study needed to answer the research questions, specifying data collection needed, and sample size and type. The **analysis and reporting** process includes developing a data presentation and analysis plan and interpreting data to draw conclusions.

Identify Topics

When confronting complex issues such as student learning and adult literacy program management, the amount of information or data you have may seem overwhelming. You can quickly become lost or confused without something to guide and focus your efforts. Having specific issues or a framework as you begin will make your efforts to use data more efficient and successful.

Exhibit 1. Evaluation Research Planning Model



Myths

Adult education and literacy programs are not well known by those outside of the field. If they think about it all, people often think any class an adult takes—such as a college or self-improvement class, “night” school, adult high school—is adult education. Lack of understanding of adult education students is also common, especially their wide diversity, needs and goals. Even those who work in education, workforce and social service programs may not have a clear or accurate picture of adult education. Adult educators find they often have to begin a conversation about their work with an explanation of the program and what it does.

Misinformation about adult education is also widespread. For example, people may think of adult education as only for high school dropouts, the students are unmotivated or the program is generally ineffective. Exhibit 2 lists some of this misinformation or “myths” and you probably have your own favorite myths that you would like to refute. Such myths are a good source of topics that you can use to begin your evaluation research and are the inspiration for the Myth Busters training and guide.

Add Exhibit 2 here

Myth Busters begins with a focus on these misperceptions about adult education as a source for ideas but there are many other sources for ideas. Performance requirements, research or knowledge of good program practice, and simple curiosity are other good ways to identify issues and topics to study.

Performance Requirements

Under the NRS, programs must meet performance standards for their student outcomes. In addition, many programs must meet other requirements set by the state or other funding agencies as part of their grant. For example, programs may have recruitment targets, be required to serve certain types of students, keep students enrolled for a certain amount of time, or pre- and post-test a required percentage of students. Many states also have instructional standards or quality indicators to promote program quality that your program must address. All of these performance requirements are benchmarks by which to evaluate your program using data. They offer areas for you to investigate that will help you manage and improve the quality of your program. You can examine whether you are meeting the requirements and variations by sites, classes, or types of students.

Research and Practices in Your Program or State

Your knowledge and beliefs about what is good practice in adult education regarding instruction, retention, intake, goal setting, and other areas can also guide you in deciding what research to conduct. We all have our own ideas, based on our experience and education, about what is important to program management. You can use data to test these ideas. The research literature also suggests other topics you might want to investigate with your program’s data, such as a new way to recruit or retain students, a revised curriculum found to be effective, or a new instructional approach. Alternately, you may suspect that something is not working right in your program and want to investigate. For example, you might know you are not retaining students long enough, not reaching your target population, or not meeting performance standards. Any of these issues can serve as the basis for study using your program’s data.

In addition, you may wonder whether practices and interventions within your program or state are effective. For example, your state may have an open enrollment policy, mandate specific instructional approaches, or require professional development workshops. There may be good reasons why your state has adopted these practices but they may not have been evaluated to verify how well they work and for which students or teachers. Likewise, your state or program may be interested in trying something new and comparing it to current practice. Online courses, for example, are increasingly popular. Are they as (or more) effective than classroom-based instruction? Thinking about specific program practices and whether they work will generate ideas for evaluation research.

Curiosity

Your interest in data may stem from simple curiosity—and this approach is not to be discouraged, as many important findings in all fields began just that way. You may want to look at data to see what you can discover. Indeed, this practice is common among people who appreciate and get excited about data. When exploring data, you are bound to find patterns that indicate good practice or problems or are otherwise worthy of further investigation.

Develop Measurable Questions

Regardless of how you decide on issues or topics to pursue, it is essential that you focus and refine them in a way that allows you to translate them into measurable questions you can address with your data. The development of good data questions is central to conducting research because the questions determine the data you will need, the research design, and the conclusions you will be able to draw. A poorly developed question cannot be answered or will provide an answer that is not helpful. Therefore, at the start of your research plans, you should carefully consider and refine the questions you are asking.

Most of the questions we ask in adult literacy address enrollment, student learning gains, retention, teachers, and instruction. It is not difficult to generate questions about these issues. A few minutes brainstorming on your own or with teachers and staff is likely to produce a long list. You may find that your list includes questions that are too broad, conceptually cloudy, or not tied to data that you can examine. There is an art to refining questions to make them useful for research and this skill improves with experience. The following information will guide you to focus the research questions, identify the data variables (inputs and outputs), ensure that your question matches the answer you are looking for, determine your data needs and sources, and plan for data analysis and reporting.

Focus the Question

A common mistake is to pose a research question far too broadly for you to answer. Although your question may address a topic of great interest, it may not be manageable because you do not have the time or resources to address it. Your question is too broad when:

- It is not one question about a topic but has several questions or topics embedded within it.
- You need to collect too much additional data beyond what you currently collect to answer it.
- It would take several years of data collection to answer it.

For example, a question such as “What is the effect of attending our program on our students’ lives?” is certainly a fascinating and important question, but it is unlikely that you would be able to answer it with your program’s data. First, there are clearly several topics and questions embedded in it. The question does not specify which features of the program might affect students' lives or what about students' lives you would expect to change. A question this broad requires a major research study and years to answer adequately. Furthermore, a satisfactory answer would require data about students' lives that you do not collect for the NRS, meaning you would have to devise additional measures and data collection procedures. Finally, it may take months or years for any effect of attendance in an adult education class to affect students' lives, more time than you have to follow up and collect data on this topic.

A broad question could be a good start, though, to begin thinking about the aspects of the program that are likely to affect students and what about their lives might change as a result. Then, think about the data you collect to form a more narrowly focused question. Some examples include: Does learning to speak English help English as a second language (ESL) students get jobs? Do students pass the General Educational Development (GED) tests? You *can* answer these narrower questions with your program's data.

Break Down Your Research Question: Educational Inputs and Outputs

One way to focus and improve a question for research is to break it apart to identify educational "inputs" and "outputs." Think of your program as providing educational inputs to students, such as instruction or other experiences that result in a change in student behavior and outputs measured through your data. Thinking about the question this way will help you evaluate whether it is a good question and also suggests ways you can narrow it.

For example, the question "Which of our classes help students more?" needs to be narrowed and sharpened. While the question does refer to educational inputs (classes) and outputs (help students), the next step is to define more clearly what these inputs and outputs are. For example:

- Class (inputs):
  - Hours of instruction offered per week
  - Teacher characteristics (e.g., education, experience, full-time, part-time)
  - Curriculum of instruction
- Help students (outputs):
  - Improve test scores
  - Advance an educational level
  - Obtain a GED
  - Improve attendance

The educational inputs, or what your program does, are what happens in class or in your program that might affect the student. The amount of instruction offered per week, the characteristics of the teacher, and the content of the instruction or the curriculum can affect student outcomes or outputs. Learning outputs related to attending the class could include improving on tests, advancing an educational level, or passing the GED tests. You might also want to examine student attendance as an output to determine whether regular attendance helps the student succeed.

Breaking down a question in this way helps you focus on what you really want to know and makes getting the answer more manageable. Going back to the example above about which classes help students more might be better phrased as “Do classes with more hours of instruction per week increase student test scores?” or “Does teacher training and experience help more students in the class get their GED credential?” These questions are more direct, can be related to NRS data you have available, and can help you with program decisions on how to design your classes.

Ask What You Want Answered

A third consideration that will help you develop data questions is to ensure that your question will give you the information you need. That is, the question, when answered, should help you resolve what works to improve your program. While this may seem obvious, it is not uncommon for people to begin research not really sure about what they want to know or whether they can use the information that will result. On the other hand, if you enjoy research and data analysis, you may get carried away and ask too many “nice to know” questions—which are interesting and satisfy your curiosity—but require the luxury of time and resources to answer.

After you break down questions, you are likely to generate several more questions, some of which you will be unable to address or will not be helpful to you to answer. For example, above we broke down the general question about classes into three topics: instructional hours, teachers, and curriculum. To decide which question you want to pursue, think about which will provide you with information you want to know and whether that information will help you make good decisions about your program. Knowing that more hours of instruction per week is related to improvements in your students’ test scores, for example, does not help you if you are unable to change your class schedules and instructional hours offered. Similarly, learning whether teacher characteristics are related to learner advancements will be of little use to you in managing your program if you are unable to hire the teachers you need. Quite simply, ask for the information you need and can use.

Refining Questions

Exhibit 3 shows sample questions at various stages of refinement. The questions in the first column are broad and unfocused. They address global concepts and imply a range of outcomes. Answering them fully would require collection of much data over an extended time. In addition, the issues they address – good teaching, program effectiveness – require several distinct subquestions.

In the second column, the questions have been narrowed and focused on data that are available in most programs. Program effectiveness in the first question has been defined in terms of student goal achievement, and an outcome has been added to help define good teaching and helping needy adults. These questions are more amenable to the next step, breaking them down to reveal the implied educational inputs and outputs for analysis. The third column shows questions that result from this breaking down process. The questions have been rephrased to address specific answers. These questions are unambiguous, identify specific data for review, and will produce answers that can help with program management and improvement.

Exhibit 3. Developing Questions for Data-Driven Decisions

Poor Question	Good Question	Better Question
Is my program effective for all students?	Do different types of students in my program achieve their goals?	How does attainment of a GED, entry into employment, and education gain differ by student age and ethnicity?
How long do students have to be in our program to be helped?	Does longer retention in our classes help our students learn?	How many hours of instruction do our students need to gain an educational functioning level?
What is a good teacher?	Does student learning differ by teacher?	Do students in classes taught by teachers who have more education and experience have higher test scores?
Is my program helping the most needy adults?	Are low literate students learning less in my program than other students?	Are literacy and beginning level ABE students advancing levels at the same rate as students who enter in other levels?

Determine Data Needs and Sources

After refining your research question into one that can be answered, determine what data you need for this work. Do you need additional data? If so, do you have time and resources to collect the extra data? If not, then refine your question so that the additional data are not needed.

If you have a very simple or specific question, the data you need will be readily apparent. For example, it is clear what data you need to determine whether older students attend class more hours than younger students. Most questions, however, are more complex and will benefit from a more formal deconstruction. One approach toward identifying data needs is to make a list or table of the topics and related data for each component of the question.

Exhibit 4 shows a table that is helpful for identifying data needs. The columns list the topics addressed by the question, the data available, other data that might be available in your program's database, and data that are not likely to be in your database. The table is constructed to illustrate the data needs for two questions: "Do teacher characteristics affect student learning gains?" and "Does student attendance and persistence differ by student characteristics and educational functioning level?"

Exhibit 4. Identifying Data Needs: Teachers, Students, Retention, and Learning

Topic	NRS Data in Program Database	Other Data Possibly in Program Database	Data Not Likely in Program Database
Question 1: Do teacher characteristics affect student learning gains?			
Teachers and Instruction	Attendance; teacher certification and years of experience (beginning in PY 2012)	Full-time or part-time status, age, gender, ethnicity	Teacher education, professional development received
Learning Gains	Educational functioning level, level advancement	Test scores	None
Question 2: Does student attendance and persistence differ by student characteristics and educational functioning level?			
Retention	Contact hours	Dates entered and exited, enrollment history, hours per week, class meets	None
Enrollment (Student Characteristics)	Student ethnicity, gender, age, employment status, highest grade attended (beginning in PY 2012)	Number and type of classes taken, weeks enrolled	None
Learning Gains	Educational functioning level, level advancement	Test scores	None

For Question 1, measures of student learning, including educational functioning level and level advancement, are available as NRS required measures. Many programs also record test scores in their databases, which could be used in this analysis. However, some of the teacher measures needed may not be available because they are not required by the NRS, and many states may not have detailed information about teachers in their databases. To answer this question, the researcher would have to collect the additional information from teachers.

All the data are available for Question 2, because the NRS requires data on contact hours, student demographics, highest grade completed and educational functioning level. Many programs also would have the additional data needed to address the question, such as dates when students enrolled and exited and their test scores. The total weeks enrolled could be computed using dates of classes, which are used as a retention measure, and test scores as measures of student learning gains.

Also consider the quality of data that are available to answer your research question. Your data system may have the measures you need; however, data that contains errors or missing data will not be useful for your research. In adult education, attendance records are sometimes incomplete or inaccurate and pretest and posttest data may be missing. Almost all students are pretested but often large percentages of students are not posttested. Consequently you may have too much missing data to make the analysis useful.

You also may find that the data you need to conduct your analysis is not available but you may believe that the question is of such importance that you are willing to use additional resources to collect that information. Some additional data, such as teacher characteristics, can sometimes be collected easily. Often, however, you may find you do not have the time or resources to collect data, and you will have to either drop the question or refine it so that you can answer it with the data that you do have. For example, you could change the question about teachers and learning gains by looking at specific teachers or classes. You could compare classes taught by full-time and part-time teachers or by teachers with advanced degrees and baccalaureate degrees. The refined question would then be: “Do student learning gains differ among classes taught by full-time and part-time teachers or by teachers with advanced degrees?” You could address this question by identifying classes taught by teachers with the characteristics of interest and compare student performance among them.

Data Needs to Control for Alternative Explanations

After you have identified data sources that address your research question, you should consider the many other factors that may affect the relationships you are studying. In educational settings such as adult education, the characteristics of learners, instructional approaches, and program factors affect the processes and outcomes you are studying. One of the goals of research is to be able to filter out or “control” these other factors to enable you to answer your research question as clearly as possible.

For example, to study the effect of professional development on student learning, you might pose the research question: “Do students in classes whose teachers received 40 hours of professional development on adult learning have higher test scores?” It would not be sufficient simply to compare test scores of classes in which teachers had the training with classes in which teachers did not. Many other factors besides the teachers’ participation in professional development could affect the results, including student characteristics, student attendance, and prior background of the teachers.

As the researcher, you will want to be able to conclude whether the professional development increased tests scores, not these other factors. To allow you to draw the conclusions you want, it is critical to identify as many of the other factors that will affect the study, so you can control for or rule them out. Exhibit 5 provides an example for our hypothetical research question. In this table, we have identified the main data elements you need to address the research question about the relationship between teacher professional development and student test scores. Other factors that might affect teachers’ behavior listed in the table include teacher background and experience, the quality of the professional development received, and whether the teacher actually implemented the principles of the training. Other factors that might affect student test scores are student background and literacy level, the level of the class, attendance, and motivation.

Exhibit 5. Identifying Data Needs: Alternative Explanations for PD Findings

Data Needs	Data Focused on Research Questions	Other Factors That Can Affect Findings	Ways to Control for Other Factors
Do students in classes whose teachers have 40 hours of professional development on adult learning have higher test scores?”			
Teachers variables	Whether teacher participated in professional development; Number of hours of participation	Teacher experience and background; Quality of professional development; Content of instruction; Teacher implements principles taught	Study teachers with similar background; Match teachers by background; Observe teachers

Student variables	Student pretest and posttest scores	Student background;
		Student literacy level;
		Class level;
		Student attendance and motivation
			Study students with similar background; Match students by background;
			Use classes at same levels;
			Statistically control for attendance and student demographics

After you have identified these other factors, your next step is to plan how you might control them so that you can rule them out as a reason for your findings. The two basic ways to do this are through your analysis approach and research design. In the analysis, you can disaggregate your data to show results separately for different types of students and teachers. For example, you could include only teachers and students with similar background on the tables and analyses. There are also ways to control for these other factors through statistical methods. Such approaches are common in research studies and are used with secondary data analyses or when resources are limited for data collection.

A second way to help rule out and control for variables is through research designs and methods. For example, you could match teachers and students on specific characteristics you think may influence results or conduct an experimental (randomized) research study.

Research Design and Methods

There are three main types of evaluation studies that you are likely to conduct. The right type for you to use will depend on your research goals:

- Exploratory study: “I want to learn more about adult education in my state.”
- Formative evaluation: “I want to know how well my pilot intervention is working.”
- Summative evaluation: “I want to know if my established intervention worked.”

The following sections provide a more in-depth look at each of these types of studies and about the type of the research method you may use when conducting your study. A final section addresses how large a study sample is needed, depending on the research goals.

Exploratory Study

An exploratory study is used before you have a specific activity or intervention to evaluate, when you are at the “just curious” stage of your research. Typically, researchers conduct exploratory studies because they want to find out either: (1) what the important factors are to consider in creating or implementing a new intervention (e.g., Is there a problem that needs to be addressed that you suspect may be more prevalent among certain types of students?), or (2) what existing intervention characteristics show promise for inclusion in a new intervention. This type of study can be guided by previous experience or theory, or it can be used to cast a wide net to uncover emergent patterns and themes in the data. The findings can then be used to improve existing theories, research, and practices in adult education.

For example, you may want to find out which student factors (e.g., literacy level) relate to attendance and persistence. This knowledge may be useful in tailoring interventions to students with differing backgrounds. Before you can begin, however, you need to decide *how* to study this issue. There are two primary methods for conducting exploratory studies—descriptive and qualitative. These two types of exploratory studies differ in focus and in the type of data used.

Descriptive study. A descriptive study has a broad yet shallow focus. It collapses a large amount of data down into a snapshot on the issue being studied. The data used are quantitative, generally extracted from existing databases and program records (e.g., attendance and posttest scores). Another common data source is surveys, but responses may need to be kept simple or converted into some type of numeric form (e.g., yes/no becomes 1/0) for analytic purposes due to the volume of data to be processed. These data can then be used to:

- Describe: “Literacy students attended 60 percent of class days.”
- Compare: “Literacy students attended fewer hours (50 hours), on average, when compared to all other types of students combined (65 hours).”
- Relate: “As hours of attendance go up, so do posttest scores.”

This last example represents a *correlational* approach. Correlational studies are a subset of descriptive studies and focus on the relationships between the factors being investigated. They *cannot* be used to infer a causal relationship, but they are useful for determining how strong a relationship is between two variables (e.g., literacy level and attendance) and what that relationship looks like (e.g., as literacy level goes up, so does attendance).

Qualitative study. A qualitative study has a narrow, but deep focus. Whereas descriptive studies rely on the type of data (quantitative) that allows for summarizing with numbers, qualitative studies collect information—often in the form of spoken or written words—on the “qualities” of something (e.g., the feasibility of implementing a new intervention) in order to develop a deep understanding. Qualitative data generally come from the following sources:

- Case studies
- Focus groups

- Observations
- Interviews
- Document reviews

Qualitative data can be summarized based on themes or patterns that emerge across the various sources of information. An exploratory qualitative study would be useful when you find that certain programs are doing either particularly well, or very poorly, and you want to understand why. By selecting a small number of programs for deep study, you can uncover the themes common to those doing well and those doing poorly, and understand the context of how adult education is being delivered in those types of programs. Because qualitative research is highly contextualized, however, your findings will not be generalizable. In other words, the scope of what you learn is specific to the programs studied. In addition, similar to correlational research, qualitative research **cannot** be used to establish causal relationships. We will talk more about why this is the case when we discuss experimental methods.

Formative Evaluation

A formative evaluation is an informal pilot test of a specific intervention, designed to provide you with feedback on how well it is being implemented and whether early signs suggest it is having the intended effect. Information learned through this type of study is used for the purposes of fine tuning and improving either the implementation of the intervention, the intervention itself, or both. After refinements are made, additional testing can be done as part of an iterative development process.

A formative evaluation consists of two somewhat independent but related studies—an implementation study and a progress evaluation. The two types of studies serve different purposes and are both important sources of information. Ideally, *both* would be included as part of a comprehensive evaluation.

Implementation study. An implementation study is useful for determining whether the intervention is being implemented as intended. It provides valuable information about the types of changes that need to be made to make the intervention more "feasible" to implement and can be useful for helping researchers identify ways to make the intervention more effective. Specifically, an implementation study can tell you:

- If each component of the intervention is being implemented (yes/no).
- If each component of the intervention is being implemented and received in the right amount, in comparison to the planned amount (right quantity?).
- If the quality of what is being implemented meets your standards (right quality?).
- If any contextual factors are creating barriers to either the quantity or quality of implementation (feasible?).

Implementation data are usually qualitative, because the purpose is to understand the implementation context and process in order to identify problem areas. Data collection methods, therefore, include those described in the exploratory qualitative section and must be tailored to each particular intervention. Multiple data collection methods may be used to provide the bulleted information above. Other (more quantitative) types of data may also be useful. For example, the number of teachers or students participating in the intervention would be important to collect. If people are not participating in the intervention, it will not be effective.

Progress evaluation. The second component of a formative evaluation provides an indication of the “effectiveness” of an intervention. The name used to refer to this type of study varies (e.g., pilot test), but regardless of the name, each study is designed to answer the following questions:

- Does the intervention seem to be improving the targeted outcomes?
- Is everyone benefitting from the intervention, or is it only certain groups?

Information about who is benefitting from the intervention—and to what extent—can then feed into changes made to the intervention. If not everyone is experiencing improved outcomes, it may mean that a different type of intervention is needed for certain students or that they need a more intensive version of the intervention. Similarly, if all types of students seem to be benefitting, but not to a great enough extent, it may mean that all students need a more intensive intervention. However, a finding of “low effect” or no effect may also indicate that the intervention is not being implemented as designed; and this is where having implementation data is useful.

The data used to determine whether an intervention is “working” will depend on what outcomes the intervention was designed to affect—student test scores, attendance, employment status, etc. These data may be collected and analyzed using descriptive/correlational or qualitative methods, or the study may use an approach called a quasi-experimental design. Quasi-experimental studies, if implemented properly, can provide a fairly rigorous test of an intervention’s effectiveness but are not always feasible. This type of study is described in more detail in the next section; quasi-experimental studies are useful approaches for both formative and summative evaluations.

Summative Evaluation

A summative evaluation is a “formal” test of an intervention (i.e., an impact or outcomes evaluation). This type of study is conducted after an intervention is well established and is designed to provide evidence that the intervention has worked (had an impact). It may be conducted by those who developed the intervention, or by an objective third party.

A summative evaluation may consist of an outcomes evaluation alone but will ideally also include an implementation study. However, unlike a formative evaluation, summative evaluation studies are more structured and reliant on quantitative data, such as NRS participant characteristics and outcomes data. The goal is to be able to answer the following questions for a relatively large number of participants:

- Did the intervention improve the targeted outcomes (i.e., have an impact)?

- Did it impact everyone, or only certain groups?
- Was it effective in all contexts or only under certain conditions (e.g., when full implementation was achieved)?

Having data on subgroups and implementation is useful for describing where, when, and for whom the intervention is (or is not) effective, which can be valuable information for your own programs as well as other programs considering similar interventions.

Formal evaluations like these tend to be large in scope, in order to support advanced statistical analysis of the effectiveness of the intervention and to test the intervention across a range of conditions. They also rely on specialized *experimental* or *quasi-experimental* research designs and methods to rule out alternative explanations for findings about an intervention's effects.

Experimental study. An experimental study is considered the "gold standard" for testing an intervention's effectiveness. These studies generally use random assignment of participant to the intervention or control (nonintervention) group, much like a lottery drawing. Using this approach, any preexisting differences between participants that could affect the results gets spread evenly across the two groups. The groups should then be comparable, on average, in every way except for whether they are assigned to receive the intervention—and a good study will demonstrate that this assumption was met by comparing participant characteristics at the beginning of the study. If the two groups are indeed equivalent in the beginning, any differences in outcomes at the end of the study can be attributed to the intervention.

Quasi-experimental study. Because it is not always feasible to randomly assign participants, you may instead choose to use a quasi-experimental design. Quasi-experimental studies are designed to mimic random assignment, most commonly through the use of a matched control group as a proxy for a randomly assigned control group. These studies can be used to make statements about the effectiveness of an intervention, as long as it can be shown that the study controlled for other factors that could explain the results. Typically, this is done by ensuring that the intervention and matched control groups were equivalent at the beginning of the study on characteristics that could ostensibly affect the outcomes (e.g., pretest scores, NRS level, language background). Then, any statistically significant differences in outcomes between those two groups are taken as evidence that the intervention was effective. A finding of effect is not considered as conclusive as it would be if it were based on experimental study, however.

Which Research Method Do I Use?

The approach you use to conduct your study will depend on your research goals, the types of data that are available to you, and what is feasible within the study context. The first step in determining the type of study or research methods to use is to specify your research goals and questions. Do you simply want to explore your data, or do you have something specific that you want to test? Do you want data that will help you improve how an intervention is being implemented? Knowing what you want to get out of the study will help you narrow down the possible approaches to gaining that knowledge. However, you can only study something for which data are or will be available. You also are limited in the approaches you can use by the context in which data are collected. Feasibility is an important issue to address in applied research. You must ask yourself what type of data and which approaches are realistically possible, given practical constraints. Exhibit 6 is a summary of the more common approaches for each research goal to help guide your decisions.

Exhibit 6. Summary of Common Research Methods, by Goal

Goal	Type of Data	Possible Methods
Explore whether outcomes vary by different participant types	Quantitative	Descriptive study
Explore relationships	Quantitative	Correlational study
Find out how well an intervention is being implemented	Qualitative or quantitative	Implementation study using expert observation, focus groups, case study, interviews, and/or document review
Find out how well a new intervention is working	Quantitative	Correlational or quasi-experimental study
Find out if an established intervention is effective	Quantitative	Quasi-experimental or experimental study (if random assignment is feasible)

Sample Sizes

Another aspect of designing your study is determining how many participants (studnets or teachers) to include. A full discussion of the technical factors and techniques for sample size determination (called “power analysis”) is beyond the scope of this guide; however, some simple and pragmatic guidelines should be sufficient for our purposes. The important thing is to collect enough data to answer your research questions. How much is enough will be determined by the type of study you are doing and what kind of conclusions you want to be able to draw from the results. If you want to make a statement about an intervention’s effectiveness across a range of program and study types, for example, you would need a sample that is large enough to represent those programs and study types. And if you want to be able to do statistical analyses, you will need a sample that is large enough to permit a fair test of your hypotheses. Simple studies, or studies that are designed to measure something with depth rather than breadth (qualitative studies), can be much smaller.

Sample sizes for exploratory studies. Exploratory studies are common among state staff using NRS data that are already available. For these types of studies, the two options are to use all available data (i.e., a “universe” sample) or to select a subsample. Selecting a subsample will be desirable if there are data quality problems with some programs. Discuss the data elements you plan to use with your data analyst before deciding whether to sample or use all available data. He or she will know how much effort it would take to process the data for all participants and if it is feasible.

If your goal is simply to *describe* the adult learner along some dimension, and it is not desirable or feasible to use data from all programs or participants, you may instead select a random sample of participants to include in the study. Selecting a random sample ensures that the results will be representative of the larger population from which it is drawn. The number to sample depends on:

- The size of the population you want to study.
- Whether you want to make any comparisons between the groups you are describing and how big a difference you expect to see.

Sample size tables are included in the appendix and assume that you would be making comparisons between groups. Therefore, if you do not plan to do any comparisons, your sample sizes can be smaller; for simplicity, consider using the “large” difference column to determine your sample sizes. Consult with a statistical expert if you are unsure how to proceed.

If you plan to do a simple *correlational study* to explore relationships in your data, you are likely to have more data available to you than you need, and you may therefore decide to select a random sample. The number of cases to include depends on how strong you think the relationship between two variables will be—the resulting sample size can be as small as 29 for a strong relationship (a correlation, or *r*, of .50) and as large as 785 for a weak relationship (a correlation of .10). For example, if you are studying the relationship between test scores and instruction, consider how much you think test scores will change relative to the instruction (i.e., a small, medium, or large amount?). For a moderate relationship (a correlation of .30), you would need a sample of 85.[1] How do you know how strong of a relationship to expect? As a rule of thumb, demographic characteristics will have a stronger relationship to outcomes (assume a moderate relationship) than will program factors, which tend to have weak relationships to outcomes. You may also consult past research on your topic to see what other studies have found. Exhibit 7 presents sample sizes for simple correlational studies.

Exhibit 7. Sample Size Needed for Simple Correlational Studies

Expected Size of Correlation	Sample Size Required
0.10 (Small)	785
0.15	347
0.20	194
0.25	123
0.30 (Medium)	85
0.35	62
0.40	47
0.45	36
0.50 (Large)	29

A more complex correlational study might look at multiple factors simultaneously in predicting outcomes. In this case, the sample sizes are adjusted based on how many variables (e.g., student characteristics) you want to relate to the outcome. Again, the sample sizes range widely, depending on the strength of the expected relationship. This time, the statistic that we care about is called R^2. It represents the proportion of variation in the outcome variable that is explained by the predictors. If student demographics are being included as predictors, it is safe to

[1] All sample size estimates in these guidelines assume a Type I error rate (i.e., a false positive) of 0.05, a Type II error rate (i.e., a false negative) of 0.20 (power of 0.80), and a nondirectional test of the statistic.

assume that you will have at least a medium-sized R^2. To determine the sample size required for detecting R^2, researchers first convert it into a universal “effect size,” f^2, which is simply a standardized R^2 that can be used to compare effects across studies. The R^2 and f^2 are similar to the small, medium, and large correlations and differences discussed earlier. See Exhibit 8 for sample sizes based on the number of variables in your study (predictors) and the anticipated size of your effect.

Exhibit 8. Sample Size Needed for Complex Correlational Studies

Number of Predictors	$f^2 = 0.02$ (Small)	$f^2 = 0.13$ (Medium)	$f^2 = 0.26$ (Large)
2	478	67	40
3	543	76	46
4	597	84	51
5	643	91	55
6	684	97	59
7	721	103	63
8	755	108	66

The topic of sample size is less relevant to *qualitative exploratory studies*, which tend to focus on a small number of cases. If you are planning a qualitative study, your sample should be selected based on substantive reasons. That means you will select your sample based on who can give you the information that you need and include participants from any of the contexts (e.g., types of programs) that you think are relevant given your research questions. So rather than a random sample, in this case you would select a “convenience” or “purposive” sample and use data from a handpicked subset of the state’s universe sample.

Sample sizes for formative evaluations. The purpose of formative evaluations is to provide informal feedback about the implementation and effectiveness of an intervention. Because it is informal, it is up to the researcher to determine what the appropriate sample size is, based on what information is needed to answer the research questions. Just remember that the results are only relevant to the specific programs, classes, and students included in the study, so make sure that you have included the types of contexts and participants that you want to learn about.

Sometimes, researchers design the progress evaluation component of the evaluation so that it is large enough to permit simple statistical testing, such as a correlational analysis or a comparison of percents or means between the intervention and matched control group (if using a quasi-experimental design). If you plan to do any correlational analyses, refer to Exhibits 6 and 7 for sample sizes, using the numbers for a “small” or “medium” sized r or R^2. If you plan to do a comparison of percents or means, refer to the state or program-level sample size tables in the appendix and use the “small” or “medium” difference column. The level that you want to make comparisons at (overall state or program level) will determine which table you use.

Sample sizes for summative evaluations. Similar to other types of studies, sample size selection for summative evaluations require that you identify a likely “effect size” because a larger sample is needed to detect a small effect than a medium or large effect. Given that most education interventions have only a small effect, a safe assumption is that the intervention being tested will also have a small effect. However, because summative evaluations serve as formal tests of an intervention, these types of studies have more complex statistical requirements for determining the size of the sample needed to test the research questions, usually by conducting some type of regression analysis (more on this later).

It is not important for our purposes to understand the technical details; however, to keep it simple, it is important to plan for including variables that represent each of the programs (minus 1; N_p - 1) in the statistical model, as well as a variable that represents whether the participant was in the intervention group. Therefore, using a table like Exhibit 7, but with likely many more predictors—(the number of programs minus 1) + (a variable representing which group each participant is in) + (background variables used as predictors)—you can see why researchers try to keep the number of predictors they include in their models as small as possible. A user-friendly sample size calculator is available at http://www.danielsoper.com/statcalc3/calc.aspx?id=1 to get sample size calculations when the number of variables you are studying exceeds the number provided in Exhibit 7. To do this, you only need to enter in the size of the effect (small, medium, or large; see Exhibit 7) you expect to get and the number of predictors. Leave the other settings at the default value.

Note on response rates. With the exception of the NRS sample size tables provided in the appendix, the sample sizes provided in this guide do not take response rates into account. Therefore, if you use Exhibits 6 or 7, or the sample size calculator, adjust your sample size by the expected response rate on the outcome data collection:

Sample size = (calculated sample size)/(expected response rate)

For example, if your calculated sample size was 1,000 and your outcomes response rate is usually around .60 (60 percent), your sample size should be 1000/.60, or 1667.

Design Data Presentation and Analysis

After you have finalized the research question, it is time to plan for the analysis. A good way to begin analysis planning is to think about how you will present the data or findings from your research. It is easy to get lost in the analyses, so having a clear vision for what you want to say is important going into the analyses.

Here is an exercise to ensure that your analyses remain focused. Take out a piece of paper. If you had only one graph to describe your findings, what would it look like? Describe the best case scenario. Put in fake data, and label all the axes and the numbers (e.g., if a column graph, label the height of the columns). Write a phrase or sentence below the graph that could serve as the newspaper heading releasing the findings. That graph is the essential component of your analyses; make sure you are collecting or have data to create this graph. If your data are qualitative, create an interesting finding.

Looking at the graph (or qualitative finding) may provoke other questions that were or were not included in your original plan. Changing the data in the graph to the worst-case scenario may also point you to other questions. This exercise will point to other graphs, analyses, and disaggregations you may want to do.

No matter what you research questions, your analyses plan should begin with a data quality check. Then, based on what type of study you proposed, you should describe the types of analyses you plan to do along with the data you need to do them and what the final report will look like. The plan should tie your question to the specific NRS and other needed measures and describe how you will present this information. In the following section, we describe the types of data quality check and analyses for each type of research study.

Pre-Analyses Data Quality Check

No matter what kind of study design you choose, your first step in analyses will always be to begin with a data quality check using descriptive statistics. “Descriptive statistics” are measures that seek to describe the data that you have. The average and range are two good examples; the average describes one aspect of the middle of distribution (sometimes called “central tendency”) while the range describes the top and bottom ends of the distribution. Looking at these statistics will ensure the quality of your data. Outliers, typos, formatting errors, and missing data problems are all common. For example, an age of “180” should be identified, deleted, and treated as missing data with documentation of the deletion. If half the students are missing a code for attendance, you might check to make sure an extra space in the data file did not cause a shift in columns. These are the types of issues people look for when cleaning and reviewing data. The easiest way to do a basic check of these potential issues is to run frequency tables or descriptive statistics on all of your variables.

Frequency tables. The most familiar form of data presentation is the frequency table. This type of table is appropriate for categorical data (e.g., ethnicity, gender) and, in its simplest form, shows the frequency (sometimes referred to as the “N”) and percent falling into each category. You will often see frequency tables with two measures, called two-way or cross-tabulation tables.

Exhibit 9 presents examples of both forms of frequency tables. The simple frequency table shows percentage and number of students by ethnicity for a program and the two-way table shows ethnicity for each site in the program. While both tables provide the total ethnicity breakdown for the program, the two-way table gives more information about how the students are distributed across sites.

Exhibit 9. Sample Frequency Tables of Ethnicity

A. Total Ethnicity

Ethnicity	Frequency (N)	Percent
Asian	76	12%
Black (not Hispanic)	120	19%
Hispanic	202	32%
Native Hawaiian or Other Pacific Islander	6	1%
White (not Hispanic)	228	36%
Total	632	100%

B. Ethnicity by Site

	Site 1		Site 2		Site 3		Total Program	
Ethnicity	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**
Asian	37	10%	39	22%	0	0%	76	12%
Black (not Hispanic)	67	17%	53	29%	0	0%	120	19%
Hispanic	145	38%	0	0%	57	80%	202	32%
Native Hawaiian or Other Pacific Islander	0	0%	0	0%	6	8%	6	1%
White (not Hispanic)	132	35%	88	49%	8	12%	228	36%
Site Totals	381	100%	180	100%	71	100%	632	100%

This additional detail in Table B demonstrates the advantage of tables that show more than one measure and disaggregates, or breaks down the data into smaller categories. Site 1, the largest site, has an ethnic distribution reflected in the program total. Hispanic and white students are about equal in size and are the largest groups in the site. Site 2 has about half white students and no Hispanic students, while Site 3, the smallest site, has primarily Hispanic students and all of the program's Pacific Islanders.

Averages and variation. When presenting data measured on a continuous scale such as test scores, we usually report the average score, which is computed by summing all the scores and dividing by the total number of scores. This average is called the *mean* and is probably the most frequently used statistic. Means are helpful for getting a sense of how a group scores on a measure. However, the mean does not always convey an accurate sense of the real "average" or what is known as the central tendency in the data. Means are misleading when there are some numbers in the distribution that are much higher or much lower than most of the others. For example, the mean is not usually a good measure of average income because some people may have extremely high incomes and others have no income at all. If you compute the mean income for the state of Washington and include Bill Gates, for example, the result will be much higher than the true average of that state's residents.

A statistic that corrects for such extremes is the *median*, defined as the number of which half the scores fall above it and half the scores below it. Medians are the appropriate measure of the average when there are great extremes in the low or high end, or range. For determining the average Washington state income, the median would provide a more accurate picture, because Mr. Gates's income and that of other wealthy individuals would not weight the average so highly in the upward direction. Another measure of the average is the mode, which is simply the number that occurs most frequently in the distribution. This measure, however, is normally used only when information on the most common score or response is needed.

Exhibit 10 shows an example of mean, median, and mode of the average number of hours of student attendance reported for the NRS by 13 states. The exhibit has the states ranked by lowest to highest attendance hours. While the overall mean for the 13 states is 92.6 hours, the median (the number that half the states fall above and half below) is 85 hours, and the mode, the most frequent average among these states, is 45 hours. Because there is such an extreme in the states' average attendance hours, the median seems most appropriate as a measure of the central tendency.

Exhibit 10. Instructional Hours per Student for Selected States: Mean, Median, Mode, and Variance

State	Average Hours per Student
State 1	24
State 2	33
State 3	42
State 4	45
State 5	45
State 6	48
State 7	85
State 8	91
State 9	102
State 10	126
State 11	176
State 12	185
State 13	202

Mean	92.6 Hours
Median	85 Hours
Mode	45 Hours
Range:	178 Hours (202–24)
Standard Deviation:	62.0

Along with presenting the mean or median, you often will want to include a measure of the variance within the data. The variance tells you how much the measures differ from the average and from each other. It is important to present variance measures, as they can provide valuable information for program management and improvement. For example, you might want to question why attendance or test scores are highly variable in one class but not another.

The simplest and most common measure of variation is the range—the difference between the lowest and highest score. In Exhibit 10, the range of average attendance hours among the states is 178, the difference between 202 hours and 24 hours. This high variation is common with student attendance, as in any class or program some students stop attending after one or two classes, while others stay for a relatively long time. Another common measure of variation is the standard deviation, which provides a sort of average variation of measures from the mean. Most software programs can compute this measure of variance routinely. The standard deviation for the above example is high, indicating attendance hours are highly variable among states.

Missing data. You may have heard that missing data can be a significant problem in doing data analyses, and it is true for several reasons. First, as mentioned earlier, if significant portions of your respondents or participants are missing, the result will only be generalizable at best to those who remained. For example, if you have a program that shows high level completions after conducting teacher professional development focusing on retention, you might conclude that the training worked. However, if you learn that half of the students dropped out or did not take a posttest, then you do not know what would have happened if they had taken a posttest. Maybe the completion rate would have still been high, but chances are probably good that it would have been lower and maybe not even significant if all "missing data" had been included.

Another important issue to be aware of is how the program you are using codes and accounts for missing data. For example, some software programs use 999 as the code for missing data, and blank spaces are considered zeros. If you think that blanks will be left out, then having them be zeros is a big difference. Consider the following numbers:

2 __ 1 9 7 __ 5 6 4 7

If treated as missing data, the average is: 5.1
If treated as zeros, the average is: 4.1

As you run your descriptive data, be sure to check how the program you are using treats blanks or codes for missing data. After you have run these analyses and are reasonably sure your data have no major issues, you can begin to answer your research questions based on the type of study you selected. We provide the following examples of analyses for each type of study: exploratory study, formative evaluation, and summative evaluation.

Exploratory Study Analyses

If your research question falls under an exploratory study, then you will likely be focused on *exploratory analyses*. Exploratory analyses are just what they sound like: analyses that uncover patterns in the data. You want to learn more about what is happening and want to look at how the data compare on a number of different dimensions—and maybe more that you have not thought about yet. You are going exploring.

You should take everything with a grain of salt, because relationships that appear to be there may be there just by chance. In fact, the reason why you should not do many of these analyses if you are testing a program, an idea, or hypothesis (for example, that your professional development program increased student completion rates) is because the more comparisons you make, the more analyses you run, the more likely you are to find an effect or a relationship when one does not really exist, also sometimes referred to as a "spurious correlation" (relationship), a "Type I error," or just a "false positive." The more relationships you examine, the more likely you are to see differences that are not real. Exploring your data is like trying to get a fuzzy picture of what is happening, then you can design other studies that try to make that picture clearer.

We will cover three ways to do exploratory analyses on your data: graphs, correlations, and regressions. Each of these provides suggestions about themes in the data. Data from qualitative methods, such as interviews and focus groups, as well as surveys also may be analyzed as a part of exploratory analyses but are covered in the next section, formative evaluation.

Graphing to explore relationships. Exploratory analyses focus on *comparisons* and graphs and are often ideal for showing and exploring relationships. You may want to compare data over time, to other data, within subgroups, and to benchmarks. A basic understanding of data presentation is essential, because often the type of presentation will determine what you actually find—or fail to find—in the data. Here we briefly review some of the basic approaches and concepts you are likely to use in data displays. While we highlight a few approaches here, the Internet provides a wealth of data on your options (see, for example, http://www.extremepresentation.com/uploads/images/choosing_a_good_chart.jpg. The Making Data Meaningful publications by the United Nations are also helpful: http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part2_English.pdf).

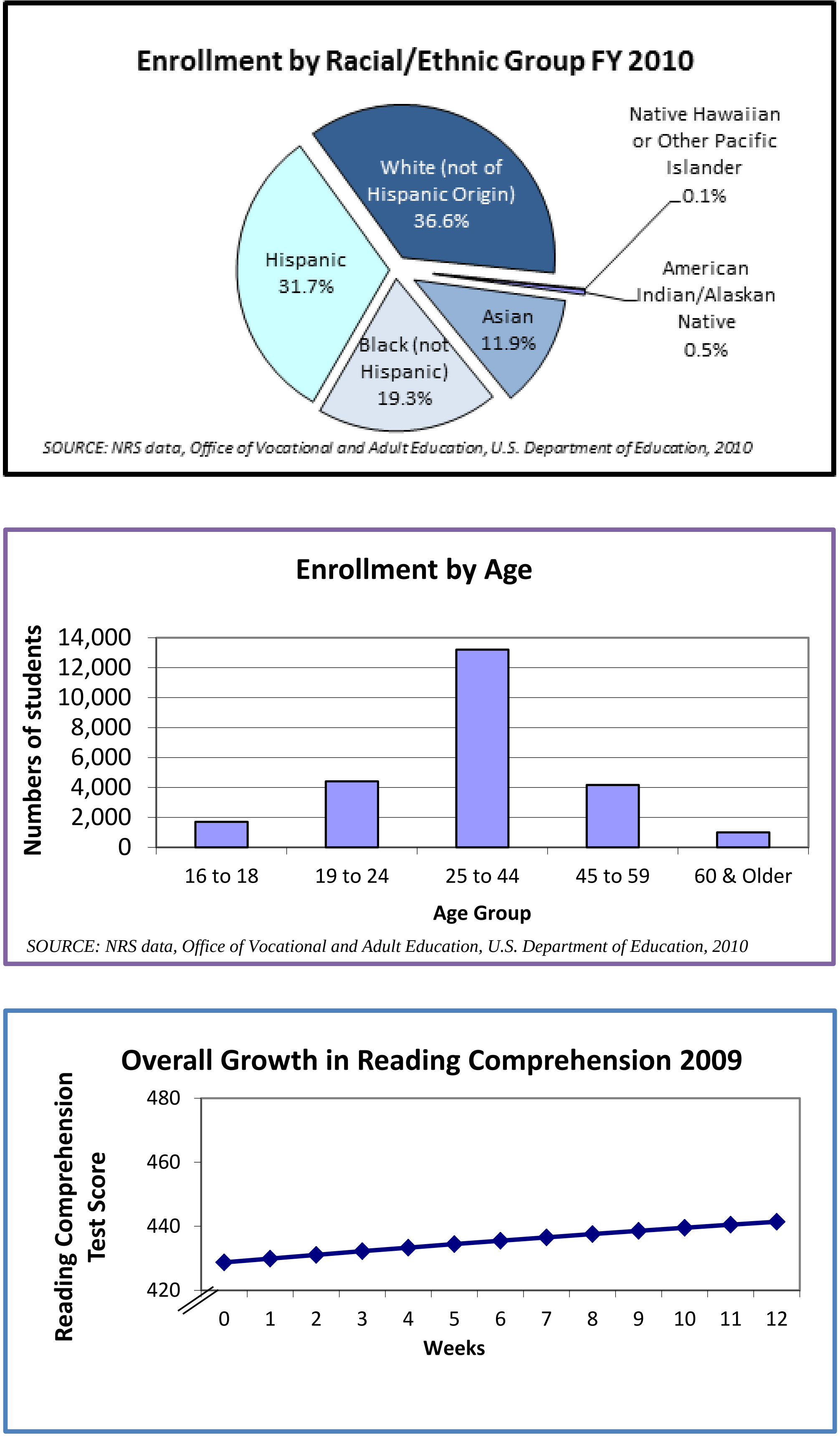
It is easy to get overwhelmed looking at rows of numbers in frequency tables. The important relationships may not stand out easily; they may be lost in a sea of numbers. Graphic presentations show data more clearly and often have a dramatic impact when showing large effects or important findings in your data. There are many different ways to present data graphically, but we will focus on the most commonly used: pie charts, bar charts, and line charts, as illustrated in Exhibit 11.

The pie chart shows program ethnicity data from Exhibit 9 in graphic form. Each ethnic group is shown as a slice of the pie, with size proportional to its overall percentage of the whole. The chart shows the preponderance of non-Hispanic white students and Hispanic students in the program.

The bar chart shows the number of students enrolled by age in the state. The categories in this type of chart are displayed as bars—the height is determined by the overall frequency or number of students in the category. This chart clearly shows that the number of 25- to 44-year-old students enrolled in the state is much greater than that of the other age groups.

Line charts, such as the one at the bottom of Exhibit 11, are appropriate when the data being studied are continuous measures (i.e., not categories), such as age, test scores or time. The example here plots pretest and posttest scores over time on a reading comprehension test. Students were tested shortly after intake and were given the posttest 12 weeks later. The average pretest score on the test was about 430 and the posttest average was 440. The chart shows the average student growth on this measure over the 12 weeks.

Exhibit 11. Examples of Graphic Data Displays: Pie Chart, Bar Chart, Line Chart



Enrollment by Racial/Ethnic Group FY 2010

SOURCE: NRS data, Office of Vocational and Adult Education, U.S. Department of Education, 2010

Enrollment by Age

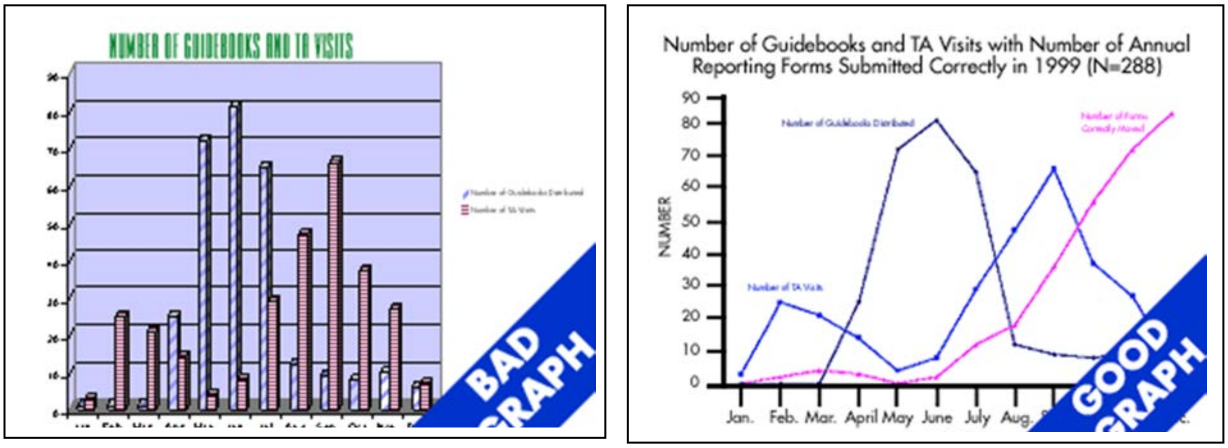
SOURCE: NRS data, Office of Vocational and Adult Education, U.S. Department of Education, 2010

Overall Growth in Reading Comprehension 2009

Some graphing tips appear in the box **Tips for Designing Data Displays**. An example of the application of some of these recommendations appears in the two charts in Exhibit 12. In the first example (left), a program manager is showing her manager the team's efforts over the last year to increase the number of annual reporting forms submitted correctly. Because the program has repeatedly had sites send the forms in filled out incorrectly, the manager decided to write a guidebook on how to fill out the forms and then send her staff on site visits to give technical assistance (TA) on filling out the forms. The first graph, labeled the "Bad Graph," shows only the number of guidebooks distributed and the number of TA visits. The graph shows that the number of guidebooks peaked earlier than the number of site visits, but that is all it shows. In the "Good Graph," you see that the number of forms submitted goes up a little with the guidebook but increases the most with the addition of the TA visits.

Additional ways that the second graph is better include: (1) the line graph (good graph) is clearer and does not use needless graphics such as the different and distracting lines in the bars of the bad graph, (2) the good graph's title gives the timeframe and number of data points contained in the graph, (3) the good graph labels the axes, (4) the good graph labels the lines directly instead of abstracting to a legend, and (5) the good graph avoids other extraneous decorations by avoiding a three-dimensional look and not using tick marks or a gray background that might distract the reader.

Exhibit 12. Graphing Techniques



As you consider what you will graph, you might consider these relationships you may want to examine and the related analyses you may want to do:

- Over time (line graph)
- Compare to others (column or bar graph)
- Compare to benchmarks (column or bar graph)
- Are subgroups very different (and hidden by averages)? (column or pie graphs, scatterplots)
- Are there any groups excelling? (line graphs of subgroups)

Many of these analyses may be done through good graphing of simple trend lines and column graphs.

Two additional analyses that you may want to run as exploratory analyses are correlations and regressions.

Correlation. A correlation is a measure of how two variables relate or co-vary. That is, when you know one of the variables, you have some information about the other. The classic example is the correlation between height and weight. Chances are, the taller you are the more you weigh. The correlation coefficient, often denoted as *r*, measures the degree of that relationship with a score between 0 and 1. A correlation of 0 means that knowing one variable tells you nothing about the other (they are completely unrelated), and a correlation of 1 would mean that you could perfectly predict the score of the second variable if you knew the first. Most people use statistical software (including Excel’s analysis add-on) to compute correlations. The more you do correlations, the better feel you will get for what is a high versus low correlation.

Tips for Designing Data Displays*

Every graph or chart should be a comparison; do not just present one group or one element. Make comparisons! They enrich context and understanding.

Don’t exaggerate; you lose the trust of your audience (e.g., although the data reports a 20 percent increase in test sores, the accompanying picture shows a picture increasing in size by 40 percent).

Use only pictures and design elements that directly relate to your point; especially avoid using purely decorative elements (e.g., does the three-dimensional component really make your point better?).

Add details if they are important (e.g., exceptions to the data), but exclude unnecessary details (such as underlining or placing text in boxes).

Make sure all labels are complete and as close to their data as possible. The title should give all necessary information (e.g., population(s) reported, dates, and sample sizes) and be able to stand alone and make sense. Begin the vertical scale at zero, unless there is a clear scale break (//), and label each axis.

Avoid legends whenever possible; label data directly (e.g., put the label directly on the pie slice of the pie graph). Don’t make the reader work to read your chart.

Find a publication that creates effective graphs and charts (e.g., the New York Times). Use these examples to design templates that fit your needs that you can use over and over.

Use color and similar elements (e.g., cross-hatching) to highlight. Don't use all colors in the spectrum on one graph; for example, use shades of blue for one group and shades of green for another.

All graphs should have source notes; from where did the data come? If multiple sources, list all.

** From Pane, 2006, and based in part on the National Center for Education Statistics Standards for Tabular and Graphical Presentations and the work of Edward Tufte.*

Regression. Another analysis you could run to examine relationships among data is a regression. You would use a regression when you wanted to look at the relationships of many variables at once. You might want to look at all the teacher variables, for example, and see which is most highly related to student completion. You could use a regression to look at this question. Your dependent variable (the one you care about predicting) would be student completion, while your independent variables would be all the teacher variables. You could do a series of correlations, but correlations will not tell you how the teacher variables relate to each other, too.

A classic example of this is when race/ethnicity of a student is shown to correlate highly with student achievement. But, take the same data and, instead of correlations, do a regression with other student demographic and background variables. You may find that race/ethnicity is no longer significant, only family income is significant. In other words, race/ethnicity in the correlation is just picking up that race/ethnicity is related to family income, and when related simultaneously, race/ethnicity does not matter anymore and income does. Statistical packages and Excel will do regressions, but people should be aware that there are assumptions underlying the use of regressions (e.g., that variables have a normal distribution) and getting assistance from researchers familiar with the issues is advised.

Formative Evaluation Analyses

Formative evaluation analyses may include a number of exploratory analyses covered above as well as additional quantitative and qualitative analyses of new data collected for the purpose of evaluation.

While there are many qualitative methods from which to choose, a common method is the “Grounded Theory” approach. The analyst would look at, for example, the transcript from the interview or focus group; identify, name, and code categories (themes) in behaviors or events; compare those categories to find consistencies and differences; merge categories as appropriate; identify emerging categories; continue until no new categories are formed (“saturation”), and identify categories of central focus, “axial categories.” Case studies, interviews, focus group interviews, and document reviews all tend to be analyzed, at least in part, through qualitative techniques.

Quantitative techniques often support or complement the qualitative data. Evaluators may report the number of interviews who reported a particular theme or how often certain words supporting the theme were mentioned. A good focus group report will give some sense of how many people agreed to various points as well as provide a text analysis.

Observation, interviews, and document analyses also often include an analysis based on scores on some kind of rubric, a scoring sheet that is determined in advance and used to frame the assessment. A classroom observation may include a list of behaviors taught in professional development, and the observer may check how often those behaviors occur or give an overall score for implementation based on a text description of levels of implementation.

Additional quantitative analyses may be desirable, depending on the type and the size of study you are conducting. All of the methods described under Exploratory Analyses, above, and Summative Evaluation Analyses, below, can also be used in a formative evaluation.

Summative Evaluation Analyses

For those doing summative evaluations, you are testing to see if something worked, for example. This type of analysis is hypothesis testing. This is when you have a theory, change, or intervention that you want to see if it worked. Did the professional development training improve teaching? Has the pay increase led to teacher longevity in the program? These types of questions have hypotheses within them: the professional development improved teaching, and the pay increase increased longevity. To test these ideas, you will do some of the same analyses summarized earlier, with the addition of a test to see if the change was big enough that it is statistically significant or if it was likely that the difference was achieved by chance. These are often referred to as tests of statistical significance and use probability to help determine whether the finding is random or “there is something there, there.”

One of the most basic and common tests of significance is called the *t-test*. The t-test is usually used to see if two means are statistically significantly different, such as the average scores of participants in a program and nonparticipants in the program. The statistic is a ratio of the differences between those means in the numerator and the range of the distributions in the denominator. To the degree that the means are enough different to overcome the variation in the data variable, then the value will be significant and this will be reported by a probability value of less than .05 (assuming this standard probability level is being used).

What does the .05 mean? It means 5 percent: Even though these means appear different enough, you could still have had these differences by chance (i.e., it was not that the professional development program was great; the numbers happened to randomly show an effect) 5 percent of the time. The .05 means that five times out of a hundred you would find a statistically significant difference between the means even if there was none (i.e., by "chance"). All of the tests of significance refer back to this idea.

Correlations, regressions, and many other analyses use significance tests to give the analyst a measure of whether the finding was likely by chance and report "*p*" or probability values. As a rule of thumb, a probability value of less than .05, the commonly agree standard, means that there was a statistically significant difference. Of course, there may not have been a *meaningful* difference, which is a part of your interpretation of the data.

There are many other potential ways to analyze data in a summative analysis, but the assumptions are numerous, so we suggest having an evaluator or statistical analyst provide guidance regarding selection of the choice of analysis.

Analyzing and Interpreting Data

With your research questions formed, measures identified, analyses completed and charts and graphs in hand, you are now finally ready to review and analyze your results to find your answers. If you are new to data analysis, you should begin with simple data displays, such as frequency tables and simple charts. You might also find it helpful to get assistance from someone who is experienced with data if you are not. We offer a general strategy below on how to analyze data and then provide some examples for illustration.

- **Answer your question.** As you look at the data, keep your original question in mind. Make sure you have the specific data elements and categories that brought you to the data and that the data match the inputs and outputs identified through your question.
- **Look for patterns and differences.** Look for patterns that stand out and differences across categories and groups of students. Look for extremes—the highs and lows.
- **Use appropriate data and statistics.** Make sure the numbers you examine are appropriate: Do you need the median or mean? Are percentages computed correctly? Do you have measures of variation? Also be alert for categories that have small numbers of students—a small "N." For example, you might notice a large difference in test scores among groups of students, but one group has only a handful of students. Do not place much faith in small numbers.

- **Draw appropriate conclusions.** If you do find patterns that answer your question, you will want to make a conclusion that closely follows what the data indicate. It is often tempting to interpret data and draw inferences that may not be warranted. Likewise, you should consider alternative explanations for your findings—that is, other plausible explanations for the patterns you find. For example, if you find the number of Hispanic students is lower at one site than at others, you might be tempted to conclude that recruitment efforts are inadequate for that group of students. The data, however, may not necessarily support this conclusion. There may be no Hispanic students in the site's service area, for example. Similarly, poor student test performance in one class compared to others at a site may not indicate poor teaching.

  Consider other explanations but do not go beyond the data. Especially remember that correlation does not mean causation. A relationship between two variables—contact hours and test scores, for example—does not necessarily mean one is the cause of the other. One way to help you draw appropriate conclusions is to talk to your staff and staff of other programs to understand procedures and types of students that may have influenced the data.

- **Remember serendipity.** Serendipity is finding something that you were not seeking. Science is filled with stories of serendipitous findings and inventions—penicillin, X-rays, "post it" notes—that proved more useful that the original objects of study. As you examine data, keep your mind open to the unexpected. Often, in answering one question, you will find that many new questions arise. Try not to dismiss something that at first may seem unexplainable, illogical, or irrelevant.

Interpreting Data: Examples

Like learning a language or learning to drive, the best way to learn to interpret and analyze data is to actually do it, once you know the basic rules and a general strategy on how to approach it. In this section, we discuss four research examples that are based on actual studies conducted by adult education researchers using NRS data. The examples illustrate how to answer research questions and use the types of analyses and presentations we have just discussed.

Question: Do 16- to 18 year-old-students complete levels at a lower proportion than students of other ages? For many programs, the most important student outcome measures for adult education are the number and percentage of students who complete an educational level. To gain an understanding of which students complete levels, one state conducted a comparison of level completions by student characteristics. Of particular concern to the state was the completion of younger students, aged 16 through 18. These students had been enrolling in higher numbers in recent years with the goal of obtaining a GED credential. The state was concerned that these students entered with lower basic skills and had lower completion rates than the older students that programs were accustomed to serving. Program directors worried that lower completion rates of younger students would adversely affect their ability to meet performance standards.

Using NRS data, the state computed a table of the percentage of completions by level for ABE and ESL students. Students were divided into three age groups, 16–18, 19–24, and 25 and older. When program directors examined this table, shown in Exhibit 13, they were surprised that, for ABE students, the reverse of what they expected was true: The older students had lower completion rates, and the younger students had the highest rates of completion in all but one level. For example, 42 percent of the 16- to 18-year-old students completed ABE beginning literacy, 35 percent completed low intermediate ABE, 54 percent completed high intermediate ABE, and 34 percent completed low ASE. In contrast, the percentage of completions of students 25 and older in these same levels were 31 percent, 16 percent, 32 percent, and 21 percent, respectively. On the other hand, in ESL, the younger and older students completed levels at about the same rate, and students 19 to 24 years old had the most completions at the lower levels. In addition, there were few completions in the higher levels of ESL.

Exhibit 13. Percent of Students Completing Levels by Age

A. Adult Basic Education (ABE) and Adult Secondary Education (ASE) Students

Age Group	Beginning Literacy	Beginning ABE	Low Intermediate ABE	High Intermediate ABE	Low Advanced ASE
16–18	42%	17%	35%	54%	34%
19–24	34%	21%	22%	50%	21%
25 and Older	31%	15%	16%	32%	21%

B. English as a Second Language (ESL) Students

Age Group	Beginning ESL Literacy	Beginning ESL	Low Intermediate ESL	High Intermediate ESL
16–18	26%	11%	7%	5%
19–24	33%	24%	5%	0%
25 and Older	25%	18%	7%	5%

Question: Do minority ABE students complete levels in the same proportion as white students? Another state was also concerned that minority ABE students had different instructional needs and incoming skill levels than white students. Program directors wanted to look at completion level by student ethnicity.

Exhibit 14 shows that the state's concerns were justified for black and Hispanic students, who had lower percentages of completers that white students at all levels of ABE. While lower percentages of Asian students in beginning literacy completed a level, Asian students' completion rate was higher than white students' completions at the two higher levels and about the same as other levels. However, the relatively few Asian students in the state cautions against making any firm conclusions about these findings.

Exhibit 14. Percent of Students Completing Levels by Ethnicity

Ethnicity	Beginning Literacy	Beginning ABE	Low Intermediate ABE	High Intermediate ABE	Low Advanced ASE
Asian (N=976)	20%	20%	29%	100%	40%
Black (N=6,742)	27%	14%	15%	28%	15%
Hispanic (N=7,723)	31%	15%	16%	32%	18%
White (N=9,751)	43%	21%	29%	61%	30%

These examples illustrate several of the points about research and data analysis presented above. The research questions were clear and tied to the data, allowing focus on the relevant patterns in the tables. The disaggregation of the state data by student age and ethnicity allowed identification of patterns that otherwise may have been missed and, for the example just cited, the few Asian students in the state cautioned us about making too much of their different completion patterns.

It is important to note that these data do *not* tell us why we found these patterns, although we may certainly speculate. For example, younger students may complete faster because their recent school experience may help them with school-based topics and tasks in ABE. They may also have a higher level of overall education than older students. Minority students in ABE may have lower education and language skills that slow their progress. This speculation should not be used to make programmatic decisions but can help guide continued data analysis.   Discussion with staff and students and review of other data will help narrow possible reasons for the differences.

Attendance and learning gains for ESL students: **Do ESL students with more attendance hours have greater test gains than students with fewer attendance hours? Do ESL students who attend class more often have higher test gains?** The relationship of attendance to student learning is of interest to many adult educators. There is a general assumption that greater attendance will result in more learning, but there is very little research on how much instruction is needed for students to make a meaningful gain. Yet, this information would be invaluable for planning instruction and setting assessment policy.   A local director of a program that served low-level ESL students wanted to know whether there was a relationship between attendance and learning gains for his students. He also believed that students who attended more regularly had higher learning gains. The director posed the two research questions above to address this topic.

Exhibit 15 shows the two tables produced to answer these questions. Table A includes only students who took both a pretest and posttest, the oral Basic English Skills Test (BEST Plus). The table divides students into three groups, based on their total attendance hours, which ranged from 21 hours to 165 hours. Much to the program director’s surprise, the students in the three groups made about the same amount of average gain on the BEST Plus. Students who attended fewer than 50 hours had slightly lower scores to start and on the posttest, gained about nine points. Students who attended most had slightly higher pretest scores and gained about 11 points. Their attendance hours were more than double the low-attending students' attendance hours, but this translated to only two points on the BEST Plus.

However, when looking at the tests scores according to percent of classes attended (Table B of Exhibit 15), a different pattern emerged. Students who attended more than 75 percent of their classes showed a much greater improvement on the BEST than did other students, especially compared to students who attended less than 50 percent of their classes. The average gain for the students who attended more than 75 percent of their classes was 19 points, while students attending 50 percent or less of their classes gained 9 points on average.

Exhibit 15. Attendance Hours, Rate of Attendance, and Test Gains for ESL Students

A. Attendance Hours and Test Gains

Total Hours Attended	BEST Plus Pretest	BEST Plus Posttest
Less than 50 (N=157)	420	429
51–100 (N=186)	422	433
101–165 (N=147)	424	435

B. Attendance Rate and Test Gains

Percent of Classes Attended	BEST Plus Pretest	Best Plus Posttest
50% or less (N=164)	422	431
51–75% (N=171)	423	439
More than 75% (N=155)	421	450

This example illustrates that sometimes you do not find what you expect to find when looking at data. You may even find something that does not seem believable and is hard to explain. Rather than ignore such findings, you should use them as an incentive to be creative and curious in your approach to data and to continue exploring what is behind the relationships. In this case, looking at how often students attended turned out to provide more insight into the relationship between attendance and learning than did looking at total hours attended. The next steps might be to try to figure out the reason for this finding—and what to do about it.

Assessment policy: Does percentage of students posttesting vary by site and by hours of instruction? Another example, about testing and assessment policy, underscores the importance of exploring data by disaggregating it into smaller units to enhance understanding. A large northeast state has a requirement that all programs pretest and posttest at least 50 percent of all students. In reviewing her NRS report, a local program director was dismayed to find that less than 48 percent of the 2,180 students enrolled had been posttested, despite hours of training and exhortations to staff on the importance of posttesting. Before deciding what to do—such as going through another round of costly training—the director decided to look at the data. She looked at posttesting percentages by providers, which included center-based and satellite sites for both ABE and ESL. She also looked at the posttesting rates according to the number of hours of instruction students had received. She set up two tables to address this question.

Exhibit 16 shows the tables. Table A reveals that the ESL and ABE center-based sites actually exceeded the 50 percent posttesting requirement and had posttested 70 percent and 60 percent, respectively. Only the satellite sites failed to meet the requirement. The satellite ABE sites seem to be a particular problem, having the highest enrollment and the lowest posttesting percentage.

Exhibit 16. Percent of Students Pretested and Posttested by Site and Instructional Hours

A. Posttesting by Provider

Provider	Enrollment	Pretested Posttested
ESL Center-based	500	350 (70%)
ABE/GED Center-based	500	300 (60%)
ESL Satellite	460	184 (40%)
ABE/GED Satellite	720	216 (30%)
Program Total	**2,180**	**1,050 (48%)**

B. Post-testing by Instructional Hours Received

Hours of Instruction Student Received	Percent Pretested Posttested
12–29 Hours	3%
30–49 Hours	60%
50 or More Hours	93%

Table B in Exhibit 16 shows that the amount of time a student is in the program—as measured by the hours of instruction received—is strongly related to whether the student is pretested and posttested. Only three percent of students who received fewer than 30 hours of instruction were pretested and posttested compared to almost all students (93 percent) who received 50 hours or more of instruction.

Based on the data, it is clear that the posttesting problem is limited to the satellite sites, especially the ABE/GED satellite site, and to students who did not stay long enough to be posttested. The solutions suggested by these analyses are that the program should try to improve testing at the satellite sites and should work on increasing retention among students who stay fewer than 30 hours.

Summary

This overview of the Myth Busters model has described an approach for planning and conducting evaluation research. We began by illustrating the importance of developing clear and explicit research questions in designing research and offered tips about how to refine and shape them. Strong questions help you clearly identify the data needed and a road map for identifying data sources and additional data needs. With this foundation, the researcher can identify other data needed and select an appropriate research design. The different types of designs, exploratory studies and formative and summative evaluations, allow you to address different questions and have associated analytic methods that allow you to draw appropriate conclusions.

Appendix
State Sample Sizes for NRS Reporting

Number of Eligible Learners in Cohort	Minimum Sample Size Required to Detect Difference of:		
	10 Percent (Large)	5 Percent (Medium)	3 Percent (Small)
1 to 200	All	All	All
201 to 400	200	All	All
401 to 600	200	400	All
601 to 800	210	500	600
801 to 1000	225	500	710
1001 to 1200	225	550	865
1201 to 1400	225	630	990
1401 to 1600	230	660	1075
1601 to 1800	240	715	1170
1801 to 2000	245	715	1220
2001 to 2200	250	740	1300
2201 to 2400	250	755	1350
2401 to 2600	250	780	1425
2601 to 2800	250	795	1480
2801 to 3000	250	805	1515
3001 to 3200	255	835	1600
3201 to 3400	255	835	1630
3401 to 3600	260	850	1690
3601 to 3800	260	855	1700
3801 to 4000	260	885	1830
4001 to 4200	260	885	1830
4201 to 4400	260	885	1830
4401 to 4600	260	890	1840
4601 to 4800	265	910	1900
4801 to 5000	265	910	1940
5001 to 5500	265	945	2100
5501 to 6000	265	945	2100
6001 to 6500	265	945	2100
6501 to 7000	265	950	2120
7001 to 7500	270	965	2200
7501 to 8000	270	965	2210
8001 to 8500	270	975	2250
8501 to 9000	270	990	2300
9001 to 9500	270	990	2330

Number of Eligible Learners in Cohort	Minimum Sample Size Required to Detect Difference of:		
	10 Percent (Large)	5 Percent (Medium)	3 Percent (Small)
9501 to 10000	270	990	2340
10001 to 10500	270	1000	2400
10501 to 11000	270	1000	2405
11001 to 11500	270	1005	2445
11501 to 12000	270	1010	2460
12001 to 13000	270	1015	2485
13001 to 14000	270	1020	2500
14001 to 15000	270	1030	2545
15001 to 16000	270	1030	2560
16001 to 17000	275	1035	2600
17001 to 18000	275	1035	2610
18001 to 19000	275	1050	2700
19001 to 20000	275	1050	2700
20001 to 25000	275	1050	2725
25001 to 30000	275	1065	2800
30001 to 35000	275	1065	2800
35001 to 40000	275	1070	2830
40001 to 45000	275	1085	2900
45001 to 50000	275	1085	2970
50001 to 100000	275	1085	2970
100001 and up	275	1100	3000

Program Sample Sizes for Optional Program-Level Estimates

Number of Eligible Learners in Cohort	Minimum Sample Size Required to Detect Difference of:		
	10 Percent (Large)	5 Percent (Medium)	3 Percent (Small)
5	5	5	5
10	10	10	10
20	19	20	20
50	43	48	50
100	74	92	97
200	116	170	188
300	144	236	274
400	163	294	354
500	178	344	430
600	189	389	502
700	198	428	570
800	205	463	634
900	211	495	696
1000	216	524	754
1250	226	585	887
1500	233	635	1006
1750	238	675	1113
2000	242	709	1208
3000	252	804	1513
5000	261	901	1895
7500	265	958	2168
10000	268	990	2337
12500	269	1010	2452
15000	270	1023	2534
17500	271	1033	2597
20000	271	1041	2646