EMSTAC

Elementary & Middle Schools
Technical Assistance Center
www.emstac.org

EMSTAC Extra

Answering the question…..

What are some guidelines that I can use to evaluate my program or intervention?

Christine Andrews Paulsen & Don Dailey
Elementary and Middle Schools Technical Assistance Center
(EMSTAC)
1000 Thomas Jefferson St., Suite 400
Washington, DC 20007
202-944-5300
emstac@air.org
www.emstac.org

IDEAs that Work
U.S. Office of Special Education Programs

A Guide for Education Personnel: Evaluating a Program or Intervention

September 27, 2002

Written by:

Christine Andrews Paulsen, Ph.D.
Don Dailey, Ph.D.

Elementary and Middle Schools Technical Assistance Center (EMSTAC)
American Institutes for Research
1000 Thomas Jefferson Street, NW
Washington, DC 20007

202-944-5300

Table of Contents

Objective of the Guide

The purpose of this guide is to provide school, district, and state personnel with an overview of the evaluation process. The guide should help these audiences plan an evaluation. If you are among these groups, this guide will provide you with an overview of what is entailed in an evaluation and issues to be aware of when planning one. It provides steps to help you get started in the planning process, identify areas where you may need assistance, and help in finding assistance to conduct an effective evaluation tailored to your program or intervention.

Please note that this guide does not provide a comprehensive tutorial or cookbook for conducting an evaluation. Instead, we hope it will serve as a useful tool for jump-starting the planning process and an easy point of reference throughout the evaluation. The Elementary and Middle Schools Technical Assistance Center (EMSTAC) can provide technical help in conducting evaluations, and we have access to a wide range of resources to address your needs.

Why is Evaluation Important?

We often hear from school and district staff that they love their work, enjoy helping students, and are highly invested in their programs and interventions designed to help students – especially students with disabilities. What these individuals also often say is that one of the most difficult parts of their job is *evaluating* their programs and interventions. First of all, it can be a daunting task to conduct an evaluation if one has never wrangled with all the technical issues involved in an evaluation before. Often, time and resources are in short supply, and conducting an evaluation may seem like added work without a clear benefit. Moreover, evaluating one's peers and colleagues, and the program they've devoted their professional (and perhaps even person) lives to, can feel awkward and uncomfortable. Given these issues, it's no wonder that undertaking a well-designed and effective evaluation can be an anxiety-producing event!

Despite these issues, evaluations are important, and the benefits that can come from a well-designed and well-executed evaluation far outweigh any challenges.

Evaluations Keep Your Program or Intervention on Track

Often we think of evaluation research happening at the end of a program or intervention's lifespan in order to determine whether the program worked – program impact. However, an equally important function served by evaluation research is monitoring *program implementation*. Evaluations of implementation are essential because they help identify problems with program

implementation *before* the program ends, so that changes in programs or interventions can still have an impact. It doesn't do us a lot of good to talk about results of an intervention if we find out the intervention was not really in place to begin with! You may find cases where the intervention changes a good bit as schools and teachers make it fit their particular circumstances or the needs of their students. Documenting and understanding these changes are important when you start to talk about how the intervention is affecting the problem or situation it was brought in to address. Without this information it may be difficult to replicate elsewhere.

Another important issue related to implementation is good old-fashioned *quality control*. How do people feel about the quality of how services are being delivered? For example, take a program designed to improve access to computers for children with disabilities. In learning about implementation we should spend some time measuring how satisfied students, parents and teachers are with the program, whether it was responsive to student and teacher needs on a timely basis, and what the challenges have been in implementing the program. Specific questions that can be answered with implementation evaluations include:

- Is the program or intervention in place?
- Is the program or intervention reaching the people (students, teachers, parents, etc.) it is intended to assist or affect in some way?
- Is the program or intervention being implemented consistent with the way it was envisioned?
- How is it changing over time during implementation?
- What are the challenges to implementing the program?
- What appears to be working so far?
- How satisfied are people with the delivery of the intervention services? What do they like most, and what are they concerned about?
- How much does it cost to implement the program? Is it feasible?
- How do the costs relate to particular services and program quality?
- Do the benefits justify the costs?

Information gathered on implementation can help you think about how to improve program services, as well as understanding what appears to be working and making a difference.

Evaluations Tell You Whether the Program or Intervention Worked

While understanding implementation is an important issue, the most common goal for schools or districts engaged in evaluation generally is determining whether or not the intervention has improved the problem or situation it was brought in to address – *program impact*. Impact evaluations answer questions like:

- Did the program accomplish its goals?

- What are the results?
- Is the program or intervention effective in addressing the problem as intended?
- How did the problem improve?
- How did the program or intervention bring about this improvement?

Answers to these questions may be combined with information collected about implementation for either a *formative* or *summative evaluation*.

Formative and Summative Evaluations

Evaluations focused on assessing program quality, implementation, and impact to provide feedback and information for internal improvement, without external consequences, are called formative evaluations. For example, a school or classroom teacher may decide to evaluate a new intervention they are implementing for math instruction. Learning how the program is being implemented, including the challenges and strong points, can serve as useful information for improving practice, rethinking how to go about things, and identifying future action steps. This information could be especially useful when combined with information on math performance. Thus, the goal of a formative evaluation is to provide internal feedback to improve practice *while the program or intervention is in progress*, rather than waiting until the program is over and you find out (too late) that the program wasn't being implemented as intended, and didn't have the results you wanted. Formative evaluations often involve both written and informal verbal discussions about results, and can be a great tool for identifying students who need assistance on a timely basis.

Evaluation studies designed to provide information on program impact to external agencies are referred to as summative evaluations. State tests administered by states to assess performance on state standards is an example of information collected to support a summative evaluation. Summative *e*valuation findings are usually reported through formal written reports, usually coming together in a final report. Thus, evaluation research also offers the added benefit of keeping stakeholders informed and satisfied. Whether it's local, district, or state policy makers or parents of students with disabilities, lots of stakeholders will want to know whether the time and resources spent on your program were worth the effort. A well-designed, credible evaluation of your program can satisfy key stakeholders that successful programs should be continued, and that implementing similar programs may be beneficial.

Of course, the best evaluation studies are those that combine both formative and summative assessments. For example, an evaluation with the most far-reaching effects, and

greatest benefits, is the one that reports findings to external agencies, while simultaneously providing internal feedback for continuing improvement.

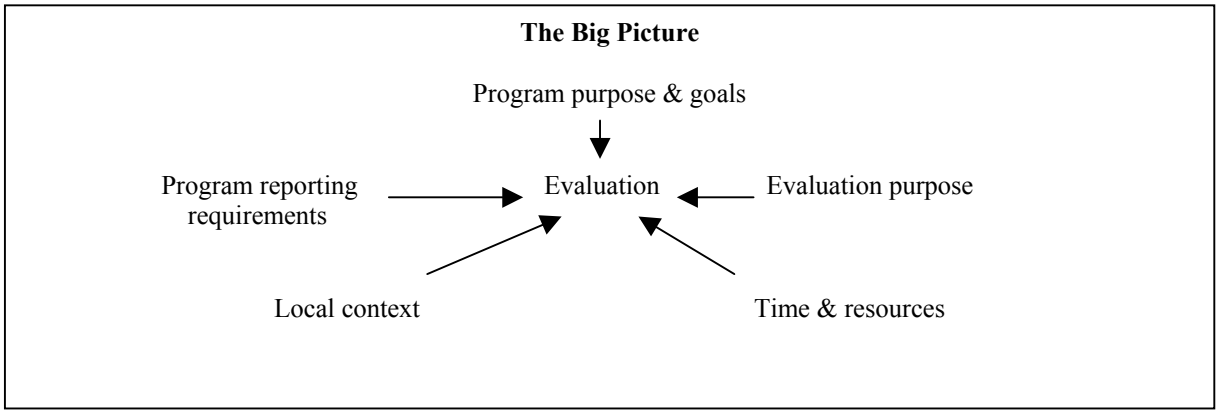
Admittedly, we are probably "preaching to the choir" in this section…If you are reading this guide you are likely aware of the benefits of a good evaluation, and just want to learn the steps involved in doing it! The remainder of this guide will walk readers through each of the important steps involved in planning an evaluation. The process begins with taking a step back and looking at the big picture. We included a flow chart diagram along each step of the way to highlight where you are in the process. The entire flow chart can be found toward the end of this product.

First Step: What Does the Big Picture Look Like?

Extra time spent during the initial planning stages of an evaluation will pay off tremendously in the end. Anticipating program changes, resource gaps, or other challenges in the beginning will prepare you to carry out an evaluation that is flexible, yet focused. While it is impossible to foresee every problem, planning ahead will help you deal more effectively with the unexpected. In fact, the very best way to plan an evaluation is to combine evaluation planning with intervention planning. This is a little easier if the program or intervention is "home-grown," meaning the school or a teacher designed the intervention themselves, but it is still entirely possible if you have adopted an intervention designed by someone outside of your school or organization. Program planning and evaluation planning can *and should* go hand in hand. Program planning provides direction for the evaluation plan. Similarly, the evaluation plan can be useful in showing how particular plans for the evaluation may need more work. When something can't be evaluated it may be a case where more concrete thinking is needed to define the intervention services or tighten up your plan of action so that people know what they are looking for.

If a program is just getting started, now's the time to put all the evaluation elements in place – even tools for collecting impact data. While you won't be studying impact for quite some time, you need to have tools (e.g., surveys, student records, etc.) in place at the beginning in order to track changes and improvements over time.

Exhibit 1: The Big Picture Issues



What are the purpose and goals of the program?

The purpose of the program or intervention will shape the evaluation. For example, the problem may be the difficulty some students experience trying to perform math computations. A school may decide to implement a program involving direct instruction that involves step-by-step sequencing and heavy use of examples to help these students improve their math performance. This intervention and its goals will likely require very different evaluation strategies than one focused on teachers working together as a collaborative planning team.

Think about how you would know whether these two different programs are being effectively implemented. One would probably involve documenting classroom activities, description of instructional activities, and review of student work. Looking at teacher collaboration would likely involve a different set of observations, and may involve more interviews with teachers individually or as a group.

What are the program reporting requirements?

Programs and interventions that receive funding or support from an external source are likely to carry with them explicit reporting requirements. Some internally-supported programs may even have reporting requirements. So, for example, a program funded by a state education agency may require schools to provide semi-annual data on school attendance or grades in a specific subject. Thus, in cases like this, some of the work has already been pre-determined for

you. Any evaluation plan needs to take into account any such requirements for reporting data to external agencies.

What is the purpose of the evaluation?

Will the evaluation be focused on providing feedback for internal improvement, or in demonstrating to outsiders that it works, or both? Make a list of evaluation purposes and goals. To find the right questions for evaluating your program, start with a long list for each evaluation goal. Ask colleagues, parents, teachers, administrators, etc. for suggestions, and then pare it back to the most important ones. This is an excellent way to expand your list of key evaluation issues, while simultaneously getting key stakeholder groups invested in the evaluation. In the end, the list should only include those things you can really find the answers to within the time and resources available.

What is the local context for the evaluation?

Evaluations do not take place in a laboratory or a vacuum, and as you have already experienced by now, there are potential political ramifications related to all the programs we implement in our schools. It will be important to think seriously about who the key stakeholders are related to your program and your evaluation. Who will pay attention to the findings? To whom will you be held accountable? Is the program controversial, or high-stakes in some way? Is the program a pet project for a prominent politician?

All of these issues are important to consider, but should not derail or prevent the evaluation process. Thinking ahead of time about the audience will help you shape the types of questions you seek to answer, how you measure important factors, and how findings will be reported.

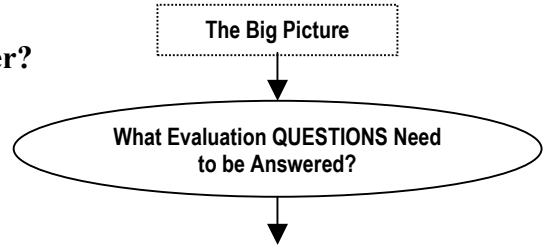
Since stakeholders consist of the individuals and groups who have an interest in the program and the evaluation, learning about their concerns early in the evaluation can provide important planning information. Stakeholders may include parents, administrators, teachers, students, support staff, community members, local, state or federal agencies. Getting stakeholder input early, and throughout, the process will also go a long way toward getting "buy-in" and gaining more acceptance of the evaluation findings.

What time and resources are available?

Think ahead about how much time you will need, how many individuals you will need to help you conduct the evaluation, and the types of help you will need. In some cases, teachers or

even parents themselves may be able to collect data. In other cases, you may require the outside assistance of an expert who can help you conduct statistical analyses of your data. Planning for anticipated needs ahead of time will enable you to re-think aspects of the evaluation, hire the necessary individuals, train personnel, or seek support from administrators.

Now that you've pondered the big picture, it's time to devise a concrete work plan.



Second Step: What Evaluation Questions Do I Need to Answer?

With the big picture in mind, you'll need to develop a list of important evaluation questions. You might find it helpful to subdivide your list into important categories, for example:

Questions related to monitoring program implementation

- Program context
- Program delivery
- Access to the program

Questions related to program impact

- Impact on student performance
- Community impact
- Impact on teacher satisfaction

How does one develop a pertinent list of research questions? Think for a moment about your favorite food. Maine Lobster? Your best friend's peanut butter cookies? Apple pie? One way to describe your favorite apple pie is to list the ingredients that go into the recipe. When you see this on paper you can begin to envision what the pie will look and taste like when finished. But, a number of things can get in the way of achieving your vision of your favorite apple pie. Unfortunately cooks vary in their ability to follow through with a recipe and instructions as laid out for them. Some will enrich and improve on things, others will fall short of your vision due to lack of experience in the kitchen. So the will and skill of the cook can make a difference. Of course, some things are beyond our control, such as the conditions we have to work with, including the quality of the stove, the quality and freshness of ingredients, and the cooking tools at your disposal. So, how the recipe and cooking instructions are implemented can lead to results that can achieve, improve upon, or fall short of your vision of the perfect apple pie.

Programs can be described in similar terms. An effective intervention will have key features and a logical story for how these features will address the problem. A good first step in looking at program implementation and impact is to make a list and develop a description of important program features and how they are supposed to work. Here you are specifying the services to be delivered. This is your recipe and cooking instructions. The program description can serve as a tool for planning how you will monitor program implementation, and give you a pretty good feel for what you should see happing in the school or classroom if the program is effective.

Evaluating program delivery and access consists of measuring whether or not the program as it is unfolding in the district, school or classroom is consistent with the program as designed and written up in your description, and how it may be changing during implementation. These issues are important because all too often the magnitude of a program's impact is sharply diminished because either the intervention is not delivered, is delivered in an ineffective manner, or it is not delivered to the right people. Knowing what took place is required in order to understand why a program did or did not work. Without this information, there is no way to determine which aspects of the intervention were effective or ineffective, or whether a larger dose or different ways of delivering the intervention would change the results. Just as cooks and kitchens differ, and this difference can make a difference in your apple pie, the abilities and resources available to schools and teachers can affect the delivery of program services. Learning how these issues may be affecting program implementation helps you identify the need for future technical assistance to improve performance.

Using the program description (list of ingredients) as a guide, you can start to develop a list of important research issues and questions for program monitoring. For example, a new program in reading instruction can use the program description to learn the following about their program:

Questions Related to Monitoring Implementation

<u>Context</u>

- Where is the program or intervention being implemented? Targeted individuals, specific classes, schoolwide, or districtwide?
- What are the important characteristics of these sites that are affecting implementation? Are there particular issues that either facilitate or constrain your ability to effectively put this program or intervention in place?
- Who are the site staff involved?

- What students are involved? How many? What particular characteristics make them eligible for the program?
- What was their performance level at the beginning of the program?

Delivery

- What instructional materials are being used? Textbook, supplementary materials, enrichment materials? Are they being used as intended?
- What procedures have been described for teachers to follow in their teaching and other communication with students? Were these procedures followed?
- In what activities were the students in the program supposed to participate? Did they participate? How did this occur?
- What activities were prescribed for other participants such as aides, parents, or tutors? Did they engage in these activities?

Access

- Are there procedures in place to ensure access to the program? For example, in a program designed for students who speak English as a second language, are teachers or other resource people available who can speak their native language?
- Do participants remain in the program as planned? When dropout rates are high, not only are the targets reached at a minimal level by the intervention, but cost per potential person may become excessive.
- Is there equal access for all potential individuals and groups for which this program is planned? For students with disabilities to participate, appropriate accommodations must be provided.

Questions Related to Program Impact

Impact on Student Performance

- Did student reading test scores improve after one year of the program?
- Were there differences in improvement depending on SES, ethnicity, LEP status, or gender?

Community Impact

- Did parents get involved in the program to the extent we expected (e.g., tutoring once per semester)?
- Were parents satisfied with the results of the program for their students?

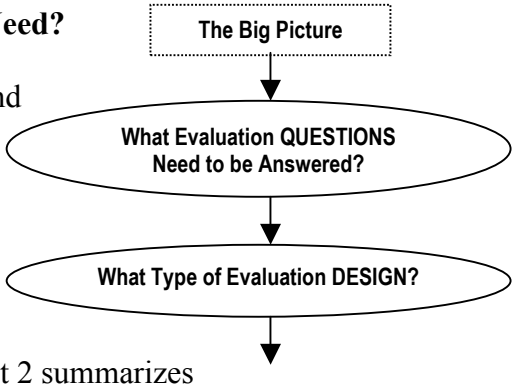
Impact on Teacher Satisfaction

- Did teachers receive the support they expected to implement the program?
- Did the program require extra preparation time, or significant curriculum modifications?
- Were teachers satisfied with the program results?
- Did the program change the way teachers delivered reading instruction in the classroom?

Next, as discussed earlier, encourage stakeholders to provide help in developing a list of important research questions. They will have valuable insight, and will feel more engaged in the

evaluation process if they believe their input is being taken seriously. In reality, time and resources will prevent you from being able to use every research question raised by each stakeholder, but this is a crucial step in the process that should not be overlooked.

Third Step: What Type of Design Will Give Me the Data I Need?



Deciding on an evaluation design can be complicated, and many of the technical issues involved in designing a study are beyond the scope of this guide. However, we provide a brief summary of different types of designs, as well as resources for obtaining more information on these designs to help you decide which direction to pursue for your own situation. Exhibit 2 summarizes the different designs discussed below. As always, EMSTAC can provide you with expert guidance in designing your evaluation.

Experimental designs tend to be rigorous in that they control for external factors and enable you to argue, with some degree of confidence, that your findings are due to the effects of the program rather than other, unrelated, factors. This control is generally achieved by randomly assigning participants. Random assignment means participants have an equal chance of being assigned to either the treatment or non-treatment group. Experiments of this nature are rarely applicable in educational settings where there is a chance that students may be denied an opportunity to participate in a program because of the evaluation design. However, there are designs that approximate experimental conditions in order to control, as best as possible, for factors unrelated to the intervention (see quasi-experimental designs).

Quasi-experimental designs are those in which participants are matched beforehand, or after the fact, using statistical methods (i.e., participants with similar characteristics are placed in both program and non-program conditions so that any differences between the programs can be attributable to program effects and not to differences between the groups themselves). These studies offer a reasonable solution for schools or districts that cannot randomly assign students to different programs, but still desire some degree of control so that they can make statistical statements about their findings.

Simple before and after studies offer a comparison of the same individuals or groups being studied at two points in time: before and after the program is implemented. A before and after design can give you some sense of impact if enough time is given for the intervention to

take root in a classroom or school. The chief problem with this design is that other factors that may be affecting performance are not being controlled (i.e., student's mature since the pre-intervention data were collected). These other factors should be taken into consideration when drawing conclusions about results. **Time series designs** are simply extended before and after studies. They offer more data points, but because there is little control over extraneous factors, and it is difficult to say with any confidence that any change in behavior over time is due to the program and not something else.

Single participant designs or case studies are those which seek to follow program implementation or impact on an individual, group, or organization, such as a school or classroom. Case studies usually provide rich in-depth descriptions that enable the evaluator to understand the important issues involved in implementing a program and the subtle aspects of its impact. Their limitation is based on the difficulty of arguing that what happens in one classroom is applicable to any other classroom. However, case studies are an excellent way to collect anecdotal evidence of program effectiveness, to increase understanding of how an intervention is working in particular settings, and to inform a larger, more rigorous study to be conducted later. While these designs do not include assignment to a control and experimental group, individuals can be compared to other students by administering norm-referenced, or norm-based assessments.

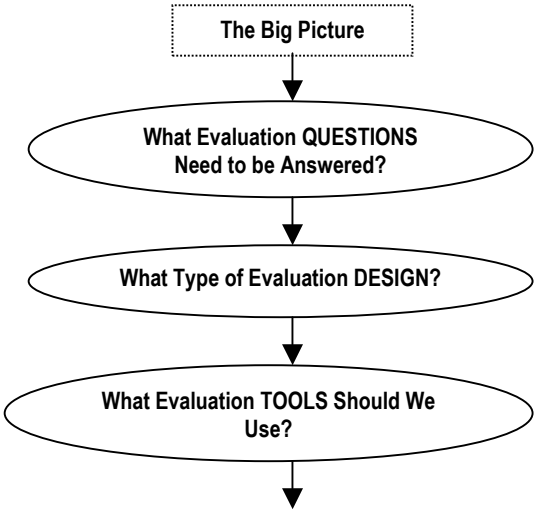
Exhibit 2: A Typology of Evaluation Designs

Research Design	Intervention Assignment	Types of Controls Used	Data Collection Strategies
Experimental design	Random assignment of participants to a group experiencing the intervention and an equal group not experiencing the intervention.	Intervention and non-intervention groups with each person having an equal chance of being selected for either group.	Collect performance data before, during, and after the intervention is implemented.
Quasi-experiments – Matched controls	Selecting schools, classrooms, or students who match on important characteristics, but are not randomly assigned.	Intervention group matched with non-intervention group on important characteristics	Before and after intervention measures of performance
Quasi-Experiments – Statistically equated controls	Participants are not assigned to a group.	Participants exposed and unexposed to the intervention are compared using	Before and after (or after only) measures of performance.

		statistical controls.	
Simple before and after studies	Participants are not assigned to a group.	No real controls	Results measured on participants before and after exposure to the intervention.
Time series – many repeated measures	Participants are not assigned to a group.	No real controls	Several repeated before and after measures.
Single participant design or case study	Participants are not assigned to a group – one individual or group closely studied.	No real controls	Several repeated over time.

In many cases a mixture of designs can work together as a design for evaluating a large, complex program. For example, the impact of a reading intervention on student performance may be compared for all students in a school over a period of time using repeated measures from exams administered for this purpose, but could also include more focused case studies of particular classes to learn about crucial implementation issues.

Fourth Step: What Tools Will Give Me the Data I Need?



Types of Tools

Once you have decided which evaluation questions are most important to answer, you'll need to determine how you are going to get the answers you need. Look at your research questions and let them help determine the "who, what, where, and when" needed to collect data. A wide variety of methods may be used (or developed) to gather data, including, for example:

- Direct observation
  - classroom activity
  - use of materials and technology
  - physical objects in hallways
  - behavior
- Records and documents
  - planning reports
  - classroom syllabi, lesson plans, grades
  - daily or weekly logs of classroom activities
  - enrollment reports
  - library records
  - test results
- Physical artifacts

    - student products
    - technology and materials
- Information from school administrators, teachers, students, and parents
    - surveys
    - interviews

Direct observation gives you a close-up view of the program as it unfolds, but also involves complications. It may be difficult to recognize what you are seeing without a well developed guide or check list on what you would expect to see if this intervention is working. Yet, even with a guide of some type it may still be difficult to determine whether or not your observation is representative or how it fits into the larger content of program delivery overtime. Records, documents, and student products can be extremely valuable in providing tangible evidence of implementation and progress, but may have similar limitations as direct observation. Well designed surveys can address some of these problems and can collect data across a wide number of respondents. Still, surveys are usually limited to collecting data about a narrow range of topics or they become too long, and they are entirely dependent on self reports made by the individuals filling out the survey, which may involve bias or perceptions that have not been well thought out. Interviews also involve self reports, but are valued for their rich descriptive detail which can help flush out understanding of how an intervention is being implemented and connecting with people. Ideally, a combination of instruments can play to the strengths of each and provide greater confidence in the information being collected. Often the best approach is one that relies on multiple sources of data to compose a comprehensive picture of what's happening in your program or school.

Spending the extra time designing a careful and comprehensive plan for evaluating your program should ease decisions about which data collection instruments to use. Some instruments are simply ready made for evaluating particular interventions. For example, classroom observations have the potential to support an evaluation of instructional interventions, and can be supplemented with interviews with the teacher and students being observed. In addition to the issues reviewed thus, other factors you will need to take into consideration include:

- Time – Some instruments take longer to develop. Longer surveys require more of a burden on respondents.
- Resources – Asking teachers to respond to a paper-and-pencil survey will require less effort than distributing and collecting a survey from parents. Conducting staff interviews

will require more resources than paper-and-pencil surveys, but may provide more in-depth insight and understanding of what is happening.

- Validity and Reliability – Validity refers to *accuracy.* How close do your instruments come to hitting the target on the bull's eye? Do your survey questions actually ask what you intend them to ask? Reliability refers to *consistency*. Regardless of where the dart hits on the target, does it hit in the same place each time? Do your questions consistently generate the same responses across people?

An evaluation planning worksheet is presented below to help you get started working through the issues discussed thus far. This worksheet could be used in planning an evaluation you will conduct yourself, and will likely expand as you keep coming back to it as a reference point and revise your planning as necessary. It could also be used to help you plan an evaluation to be conducted by an outside evaluator. Finally, it can help you be an informed consumer of an evaluation by helping you identify questions to ask and key issues to look for. Exhibit 3 provides a short sample of what a district might develop if they were interested in evaluating the success of a new reading literacy program for students with disabilities. Of course, in reality the number of evaluation objectives and questions would be greater. But, this should serve as a good starting point. Notice the range of different instruments and measures that can be employed to evaluate a program or intervention.

Exhibit 3: Evaluation Planning Worksheet

Evaluation Objectives	Evaluation Questions	Instrument	Specific Measures
A. To determine the extent to which our reading literacy program is helping fourth grade students with disabilities improve their reading skills	1. How has the students' reading test scores improved from pre-program to now?	Reading Scale	Total scale score, at two points in time
	2. In what areas are students' showing the most & least improvement?	Reading Scale	Subscale scores, at two points in time
	3. Are there differences in performance for different groups of students (ethnicity, LEP, low income) over time?	Reading Scale	Total scale scores, sorted by group, over time
B. To determine whether our reading literacy program is improving teachers' instructional skills in reading	1. Are teachers demonstrating the important skills emphasized by the program more now than before the program started?	Classroom observation	- Time on task - Attention to students with disabilities - Component checklist - Others
	2. Are teachers' overall performance ratings improved since before the program?	Performance appraisals	Overall supervisor rating of teachers' skill level

	3. Are teachers reporting that they are more comfortable with the skills required now than before?	Teacher survey	• Teacher impression of their skill level • Teacher perception of the effectiveness of program • Teacher perception of helpfulness of training/supervision under the program
C. To determine the level of parent satisfaction with our reading literacy program	1. Are parents satisfied with their children's experience with the program?	Parent interview (by phone)	• Satisfaction with child's performance in reading. • Satisfaction with teacher. • Desire to see program continue.

Developing the Tools You Need

Once you have decided what instruments and tools to use, you'll need to either collect them (if they are existing instruments), or you'll need to develop them. The first step in developing instruments is to have a copy of your evaluation plan and questions handy. The instruments should be fully aligned with the questions you need to answer – otherwise you will be wasting time and resources collecting data you will not use. The worksheet above should help you to think about what types of data you need.

The next step is to begin developing the specific items for your observation instrument, interview guide, or survey. At this stage it is always helpful to draw on guidance from different sources such as EMSTAC. In the end the best instrument is one that will give you the information you need to answer the questions guiding the evaluation.

Observation instruments

The successful observation instrument usually has two characteristics: it obtains detailed information specific to the targeted site, group, or individual (school, classroom, teacher), and it collects information allowing comparisons. A wide variety of instruments can be used for observing classrooms or other settings, ranging from a highly structured checklist of what you would expect to see if everything is going well, to simply capturing events as they unfold in a particular setting with descriptive field notes. Your choice of method should be influenced by how much is known about the program being observed, the nature of the program, and the experience of the observer.

EMSTAC is particularly sensitive to the unique demands of trying to capture the subtleties of both teacher and student behavior during classroom observations. Through our

extensive evaluation work, we have a pool of established observation protocols and experienced observers at our disposal. A sample of protocol items used to observe mathematics instruction is provided in Exhibit 4. By using a combination of observation techniques, you can acquire both rich narrative descriptions useful for putting the findings in context, and specific checklist style ratings focused on particular program elements.

Exhibit 4: Sample Classroom Observation Protocol

Cluster:_____

School name:_____

Date of visit:_____

Time arrived: _____

Time departed:_____

Observer(s):_____

1. Classroom Environment

Diagram the physical layout of the classroom. Label student desks, tables, or workstations; teacher desk; chalkboards, overhead projector, computers, and other important pieces of equipment; windows and doors.

2. Observed Learning Environment

During the observation period, keep a running log of teacher and student activities. Make a note of the time of each activity. Note important comments made by teachers and students, pieces of lesson presented, problems presented, materials used, activities presented, etc.

Time	Teacher	Students	Notes and comments

Mark the extent to which the teacher and the students used the following materials and equipment during this class period.

	Teacher				Students			
	Not observed	Observed to a very limited extent	Observed to some extent	Observed to a high degree	Not observed	Observed to a very limited extent	Observed to some extent	Observed to a high degree
Text	0	1	2	3	0	1	2	3
Workbook/Worksheets	0	1	2	3	0	1	2	3
Notebook	0	1	2	3	0	1	2	3
Chalkboard/Whiteboard/Easel	0	1	2	3	0	1	2	3
Overhead projector	0	1	2	3	0	1	2	3
Appropriate calculator	0	1	2	3	0	1	2	3
Protractor, compass, ruler, etc. (math tools)	0	1	2	3	0	1	2	3
Film or videotape equipment	0	1	2	3	0	1	2	3

Computers/software	0	1	2	3	0	1	2	3
Manipulative, models (objects used to concretely model math tasks or problems such as blocks, algebra tiles)	0	1	2	3	0	1	2	3
Other (indicate) _____	0	1	2	3	0	1	2	3

6. Did all students study the same content?

   Yes No (If no, explain)

7. Lesson vignette: Describe goals, topics, and activities of the lesson observed. Describe teacher's instructional strategies and techniques. Describe student activities.

Interviews and Surveys

Interviews and surveys can be used to collect information from those individuals participating in the program (teachers, students) to learn from them about their experiences and impressions. Interview questions can be either broad and open, highly structured, or a combination of these two formats. Open-ended questions are designed to generate longer reflections and responses from the interview respondents, and are often used to explore something about which you and the research literature have limited understanding. Highly structured questions are designed to elicit shorter and more focused answers, and are more easily structured by what you already know about the subject at hand. Given the nature of the program or intervention, some combination of open-ended and structured questions may be most appropriate.

If you are planning to use a survey, there are numerous, well-written, and thoughtful books on how to write effective survey questions. We will outline the appropriate steps to take here – but for more information on these topics, we encourage you to check out the following:

- Schwarz, Norbert & Sudman, Seymour (Eds.). (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research.* San Francisco, CA: Jossey-Bass Publishers.

- Sudman, Seymour, Bradburn, Norman, & Schwarz, Norbert (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco, CA: Jossey-Bass Publishers.

- Belson, W. (1981). *The design and understanding of survey questions.* Aldershot, England: Gower.

Once you have compiled a list of questions to ask for either your survey or interview guide, you'll need to weed out the ineffective or just plain bad ones. One way to do this is to review the list of questions, and remove any that simply seem to have little face validity – in other words, they simply don't seem like they will get the information you want. It may seem superficial, but if you question the ability of a particular survey item to get the data you need, one of your stakeholders probably will, too. Face validity is about strengthening your credibility.

Steps to Develop Interview and Survey Questions

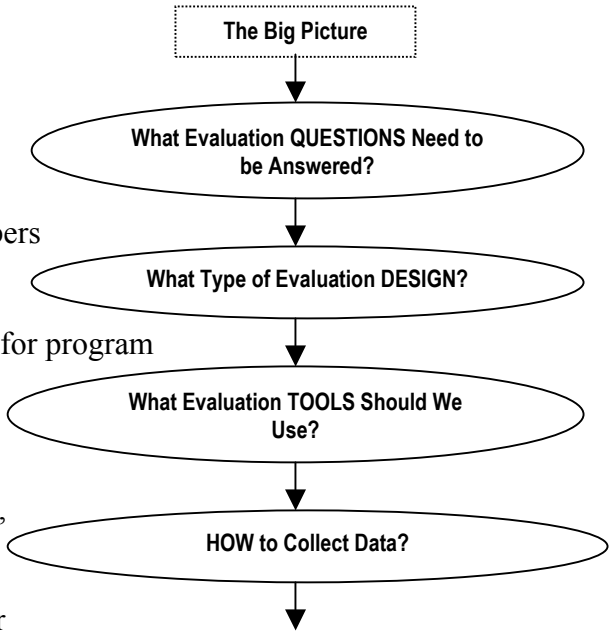
- Develop potential questions
- Clean questions
- Pilot test
- Debrief potential respondents

The next step is to conduct some small-scale pilot testing. This can be as formal or as informal as you like, but it is an important step that can save you lots of time and effort in the end. Pilot testing consists of administering the questions to a small group of potential participants. After the participants have completed the survey or interview, you should debrief them with the following types of sample questions:

- Were there any questions that were confusing to you?
- Did you understand what was expected of you for Question 3? Please explain.
- When you answered question 4, how did you know which choice to pick?

These are just a few examples of potential debriefing questions. The objective of the debriefing, and the pilot test itself, is to *make sure that the questions you are asking are the ones you intended to ask*. Participants in a pilot test may reveal that they thought you were asking one thing, when you were really trying to get at something else. So, pilot testing and debriefing with potential respondents is really the only way to determine whether your questions are going to enable you to collect the data you need. This may seem trivial, but these issues can be especially important in developing surveys. A poorly worded survey question, or one that contains mistakes or typos, can render a survey question useless. So, pay special attention to the fine details.

Depending on the nature of the program or intervention, one particular method of data collection may stand out as especially effective in helping you get the data needed to answer your evaluation questions, though often using a combination of observation, interviews or surveys, and review of important documents will give you the most confidence in your results. These methods have varying strengths that bring different issues to light, and results from one instrument can confirm those found from another.



Fifth Step: How Do I Collect the Data?

Many schools and districts do not have the luxury of having an evaluation department with a number of staff members dedicated to the purpose of collecting data and administering data collection instruments. In many cases, the responsibility for program evaluation falls to one key individual. Depending on your timeline, and the immediacy of your needs, it may be necessary to draw on other staff members, outside consultants, or even parents, to assist with collecting data.

Some schools have enlisted teachers to help administer student assessments or tests. Others have asked parent volunteers to mail surveys to families. Still others have asked for help from the various technical assistance centers, consulting firms, universities, and other agencies around the country.

Training Data Collectors

Whether you rely on internal staff to help, or employ the services of an external party, the data collectors will need training. Plan to spend at least a day training data collectors for each of the instruments they will be using. A typical training might cover the following:

- **Orientation to the evaluation**
  - Including a discussion of the evaluation objectives and research questions
- **Detailed explanation of the data collection instrument(s)**
  - Walk through each and every item on the instrument and explain its purpose
- **Demonstration of how to use the instrument**
  - For a survey or focus group protocol, this demonstration could be a role-play; for an observation tool, the trainer might show a videotape of a classroom, and demonstrate how to rate the behavior
- **Appropriate behavior**
  - Discuss how to behave professionally and appropriately when interviewing or observing other people
  - Interviewing techniques
  - Confidentiality and protection of human participants
    - Explain the rights participants have when involved in any research study, and how to avoid breaching them

- **Practice, practice, practice**
  - A significant portion of the training time should be spent using supervised practice in which data collectors try out the tools, and receive immediate feedback on how they are doing

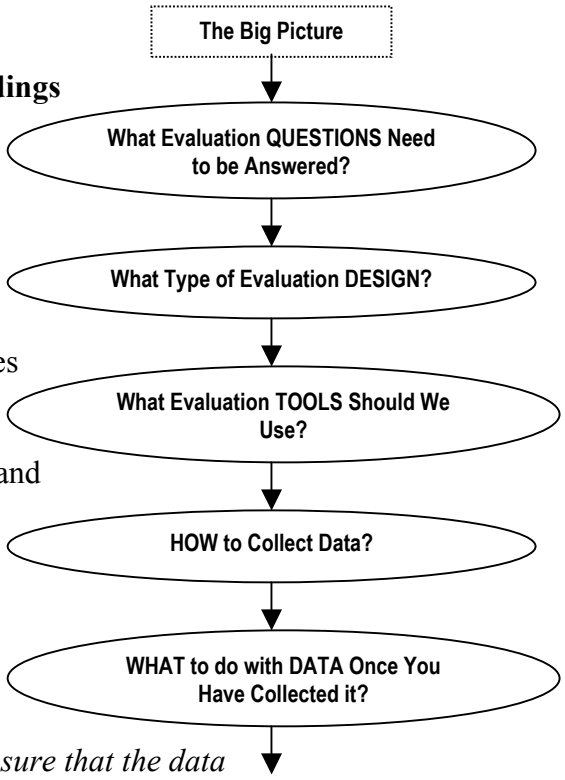
Training should occur in advance of data collection, and the two should occur as close together as possible. This will help to ensure that data collectors remember everything they learned in training. If resources allow, data collectors should be periodically observed in the "field" as they are collecting data, to ensure they are receiving the supervision and additional training they may need.

Permission to Collect Data and Informed Consent

Any time that individuals are included in an evaluation study, there are potential risks to them that can result from their involvement. The risks can range from minimal to serious, but the potential for risks should never be ignored. Students could be negatively impacted if their test scores were accidentally released to a third party, teachers could suffer serious consequences at work if their survey data were shared with administrators, etc.

In fact, if your evaluation is funded in part by a government agency or university, there are typically Human Subjects Committees and Internal Review Boards (IRBs) that will need to review your work plan and require you to answer some very detailed questions about your evaluation plans before proceeding. These boards are established for the purpose of protecting study participants. At the very least, they will require you to obtain permission from study participants and, in the case of students, will generally expect you to obtain informed consent from students and their parents. Regardless of whether you have funding from an external source, it is standard ethical practice to obtain permission and informed consent from participants before a study begins.[1]

[1] For more information on ethics and standards, see the guide written by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education: *Standards for Educational and Psychological Testing* (1995). Washington, DC: Authors.



Sixth Step: How Do I Analyze the Data and Make the Findings Useful?

This section will provide some guidance on how to prepare and implement an analysis of your evaluation data.

It is not intended to provide an exhaustive discussion of statistical or qualitative analysis. There are literally volumes written on this subject. The information provided in this section should enable you to determine what you need and how to get it.

Storing Data

Depending on the space you have, storing the data may or may not be a concern. Regardless of where you keep the surveys, forms, and datasheets, *you should always ensure that the data are kept safe and confidential*.

A locked filing cabinet will usually suffice. Security of the data is especially important in the case of test scores, or surveys that have peoples' names on them. In order to get people to agree to participate, you promised their responses would be safeguarded. You must respect that or risk losing people's participation in the next round of data collection.

To prepare for analyzing the data, it may be most efficient to transfer the data from paper to an electronic database. There are several database types on the market – some more complicated than others. It is often unnecessary to learn how to use technically sophisticated software for storing most data. For quantitative data (numbers or short text), simple spreadsheets will do (prepared in a software package like Microsoft Excel®). These can be easily imported into a statistical package like SAS® or SPSS®. Exhibit 5 provides a sample spreadsheet containing data from a parent survey.

The first column represents the respondent's confidential ID number (which can be matched with a separate list of names and corresponding ID numbers), so names do not have to appear in the database. This protects the respondents' identities. The second column represents the child's grade level. The third column represents the child's reading teacher. The fourth column represents whether the student is classified as special education. These can also be entered as numeric values, as we'll see in the other columns. Statistical packages like SAS and

SPSS don't care which way you do it, as long as you tell them what the numbers mean. The fifth column represents the answer to the question: Are you satisfied with your child's progress in reading this year? A 1 means Yes, 0 means No. The sixth column represents the answer to the question: Would you like your child to continue learning to read under this program? Again, 1 means Yes, 0 means No. You can create practically an infinite number of columns and rows for each database.

Exhibit 5: Sample Spreadsheet

	Col 1	**Col 2**	**Col 3**	**Col 4**	**Col 5**	**Col 6**	**…so on**
Row 1	**ID**	**Grade**	**Teacher**	**SpecialEd**	**Satisfied**	**Future**	…
Row 2	1002	4	Brown	Yes	1	0	…
Row 3	1003	4	Groves	No	0	1	…
…so on	…	…	…	…	…	…	…

For data that are heavy with text, like open-ended responses to survey or test questions, you may want to consider a software package like NU*DIST®, which can perform content analyses or provide other qualitative summaries.

Analyses

There are several types of analyses that may be performed with your evaluation data. This section of the Guide will not endeavor to explain how to conduct any specific statistical test. Instead, we provide an overview of the types of analyses that are available and the questions they are designed to answer, so that you may be better informed when planning for data analysis.

Qualitative Analysis

The goal of qualitative analysis is to make sense of the collected data in ways that capitalize on continuing refinement and ensuring maximum understanding of the concepts and relationships being studied. Qualitative analysis is conducted with data that are not easily represented by numbers. For example, anecdotal experiences, stories, observational data, and ethnographic data, to name a few, are best analyzed and presented in textual form, rather than attempting to reduce the observations to numerical data and subjecting them to statistical analysis. Qualitative analyses are often designed to corroborate findings from quantitative analyses, identify new leads, and provide close up examples of behavior and practices pertinent to the evaluation.

While there is no one right way to analyze qualitative data. The following steps will provide a sense of how this can occur:

- Develop codes that will help you organize and analyze collected data;
- Check and clean collected data;
- Organize collected data by people, places, or topics;
- Review data as organized to identify patterns consistent with your codes;
- Revise your codes based on an initial review of the data and emerging patterns, followed by a second review of the data in light of the new codes;
- Connect patterns together in the form of key relationships; and
- Search for alternative explanations.

There are computer-based programs that can help you summarize and categorize the data for easier analysis and reporting. These programs, like NU*DIST and Socio-GRAM, are helpful in conducting content analyses.

Quantitative Analyses: Descriptive and Univariate Statistics

When data are represented numerically (e.g., test scores, grades, score on an attitude scale), we have several choices for analyzing and interpreting them. Descriptive statistics do just that – they describe what the data look like, without making any statements about relationships between phenomena. Descriptive analyses are an important precursor to conducting inferential statistical analyses because they will inform you of the properties of your data, and may indicate the need for a particular variation of a statistical test. Descriptive and univariate (one variable) statistics that can be computed include:

- Sum
- Range
- Mean
- Median
- Mode
- Percentile
- Standard deviation
- Standard error
- Standard scores
- Skewness

- Kurtosis
- Variance

Each of these provides a descriptive piece of information about your data. Short definitions for these statistical terms can be found in the Glossary section of this document.

Quantitative Analyses: Inferential Statistics

Inferential statistics enable you to look at two or more variables in relation to each other, and, with some degree of confidence, make statements about whether the relationship could have occurred by chance or whether the observed relationship appears to be "real." For example, say you have data on student grades in 8th grade mathematics and student scores on the math SAT. You want to look at the potential power of 8th grade math scores to predict later achievement on the math SAT. You conduct a correlation analysis and find a positive correlation that is high (in terms of magnitude) – say .80. That number is meaningless unless we examine the statistical test which tells us whether the .80 correlation might have occurred by chance. If it could have happened by chance, then we can't say anything about the relationship between 8th grade math scores and math SAT scores. The "p-value" associated with the correlation helps us determine how likely the results are due to chance. When we see p-values less than .05 (or 5%), we refer to these tests as *significant*. When a researcher says a correlation is significant at the .05 level, they mean that there is a less than 5% probability that the observed results are simply due to chance. Thus, in our example, if your .80 correlation was associated with a .05 p-value, you can state, with much confidence, that there is a positive and significant relationship between 8th grade math grades and SAT scores.

Some of the inferential statistics available to you are shown in Exhibit 6.

Exhibit 6: Inferential Statistics

Test	What it Tests	Example Question
t-test for independent samples	Difference between two independent groups with respect to one variable	Does Classroom A differ from Classroom B with respect to reading achievement scores?
t-test for paired samples	Difference between two correlated or paired groups with respect to one variable	Do students in School A do better on the science achievement test from Grade 9 to Grade 10?
Correlation coefficient	Strength and magnitude of the relationship between two variables	Is family income related to math test scores?
One-way analysis of variance	Difference between two or more independent groups with respect to one variable	Is there a difference between Classrooms A, B, C, and D with respect to attendance rates?
Regression analysis	Ability of one or more variables to	How well do family income, number

	predict another variable	of languages spoken at home, and parental marital status predict sharing behavior in 6 year olds?

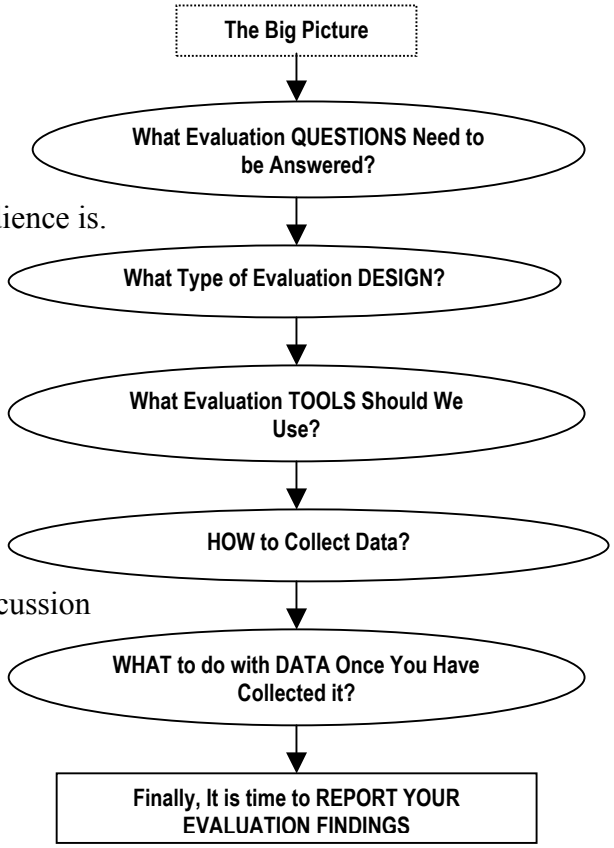
There are many variations on these tests to accommodate different types of data, number of variables being studied, sample sizes, etc. But, these represent the most common basic inferential statistical tests.

Non-Parametric Tests

Since we do not live in a perfect world, it is often the case that the studies we conduct are not perfect either – even with a well-planned and executed design. Sometimes the data simply don't want to cooperate, and they exhibit characteristics that make inferential statistics less appropriate or desirable. For instance, you may end up with far fewer participants in a particular sample than you anticipated. Or the data are spread out all over the place (e.g., you may have a bunch of people clustered at either end of a scale – half the group has an average scale score of 1, while the other half has an average scale score of 7, with no one in the middle!).

In cases like this, you may want to explore non-parametric statistics. These are a bit less fussy than their parametric (inferential) cousins, and offer an excellent option for "uncooperative" data. Some of these tests include:

- Chi square
- Mann-Whitney U test for two independent samples
- Kruskal-Wallis test for two or more independent samples
- Wilcoxon rank sum test for two correlated samples
- Spearman rank-order correlation coefficient



Reporting the Findings

How you report the data depends on who the audience is. A report written for education professionals will undoubtedly be different than a report written for dissemination to parents.

Remember to keep it simple and straightforward, no matter who the audience will be. But, also try to include information that will be of interest and use to the various audiences. A detailed discussion of the distribution of the standard errors may be appropriate for researchers, but needs to be explained in layperson's terms for parents, practitioners, and policymakers.

Most individuals have a lot on their plates these days, and will appreciate a brief report containing only the essential facts. Some policy makers prefer no more than a one-page summary of findings and implications for policy.

There are a number of technological advances that provide a multitude of options for disseminating research reports and findings. Besides hard-copy "paper" reports, findings can be posted to a website, stored on a CD-ROM, or distributed by e-mail.

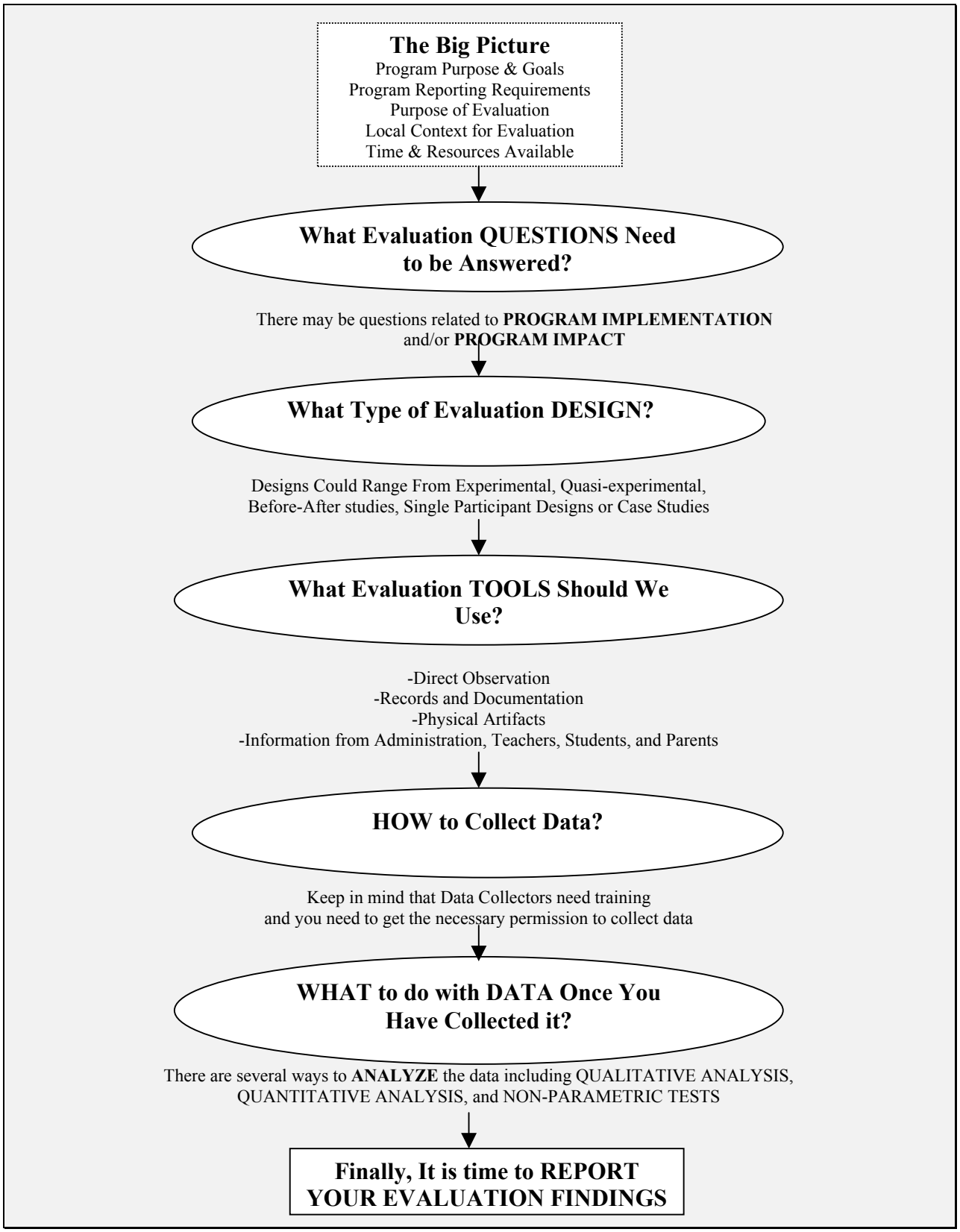
To get the attention of parents or hard-to-reach stakeholders, you may need to get a little more creative – perhaps publishing important results in local newspapers, getting some air-time on a local talk radio station, attending PTA meetings, or handing out leaflets or flyers at local businesses and libraries. These may seem to be unconventional methods of disseminating evaluation findings, but they may offer the only opportunity to tell parents that your reading programs are working for their kids.

Conclusion

We hope this guide has helped you think about the important issues to consider when planning an evaluation. You will find a logic diagram on the following page summarizing the main steps for consideration when embarking on this process. We also hope that we have encouraged you to want to learn more about evaluation, and that you will check out the resources mentioned throughout the guide, as well as our extensive list of evaluation methods and

resources that follows. And, remember that EMSTAC is available to provide input to you as you develop and implement your evaluation!

Conducting an Evaluation: A Logic Model



Evaluation Resources

Following is a short glossary of basic statistical terms and a list of evaluation resource citations. Some are more recent than others, but the older ones are considered seminal works in their field and are still used widely today. This list is followed by resources for statistical and measurement resources, and useful websites.

Glossary

- The **sum** is the total of the all scores in a particular sequence.
- The **range** is the distance between the highest and lowest score. Numerically, the range equals the highest score minus the lowest score.
- The **mean** is one of several indices of central tendency that statisticians use to indicate the point on the scale of measures where the population is centered. The mean is the average of the scores in the population. Numerically, it equals the sum of the scores divided by the number of scores. It is of interest that the mean is the one value, which, if substituted for every score in a population, would yield the same sum as the original scores, and hence it would yield the same mean. (e.g., take the sequence 7, 9, 14, and 16. The mean is calculated as (7+9+14+16)/4=11.5
- The **median** is another indices of central tendency that statisticians use to indicate the point on the scale of measures where the population is centered. The median of a population is the point that divides the distribution of scores in half. Numerically, half of the scores in a population will have values that are equal to or larger than the median and half will have values that are equal to or smaller than the median. In a sample with an odd number of observations, it is the middle observation or score: in the sequence 7, 9, 13, 17, 20, the median is 13. With an even number, the score is halfway between the two middle ones. Using the first sequence (7, 9, 14, 16), the median would be 15.
- The **mode** is simply the score, measure, or category that occurs the most often. Note that the mode is not the frequency of the most numerous score. It is the value of that score itself. Also notice that if there are two (or more) different scores that occur with equal frequency and that frequency is higher than the frequency of any of the other scores, the population is described as multi-modal.
- The **percentile** is the percentage of cases in a frequency distribution that fall below that score (e.g., the median is the 50^{th} percentile because 50 % of cases have lower scores than the median).
- **Standard deviation** and **variance** measure how the observations are spread around the mean. Standard deviation characterizes the dispersion among the measures in a given population. The standard deviation is the positive square root of the variance. To obtain it, we find the distance of each observation from the mean, square that distance, find the mean of those scores, and take the positive square root of that mean. The formula for the standard deviation (SD) of a set of scores is

$$\sqrt{\frac{\sum(\chi - M)^2}{N}}$$

where Σ indicates summation, χ stands for each score, M for mean, and N for the total number of scores. The mean is subtracted from each score; these differences are summed; the sum is divided by the number of scores. The positive square root of that quotient is the

standard deviation. The standard variation is the square root of the variance; the **variance = SD²**

- **Standard error** or the **standard error of the mean** is the standard deviation of a sampling distribution. It is an estimate of the standard deviation of the sampling distribution of means, based on the data from one or more random samples. Numerically, it is equal to the square root of the quantity obtained when **s squared** is divided by the size of the sample. The formula for the standard error is:

$$\sqrt{\frac{S^{\frac{2}{x}}}{n}} = \sqrt{S^{\frac{2}{x}}}$$

- **Standard score** is the raw score divided by its standard deviation. Also known as the **z-score.**
- **Skewness** is a deviation of the frequency distribution from symmetry. Positive skewness has a long tail to the right toward the high scores; negative skewness, the opposite.
- **Kurtosis** measures the "heaviness of the tails" of a distribution (compared to a normal distribution). Kurtosis is positive if the tails are "heavier" then for a normal distribution, and negative if the tails are "lighter" than for a normal distribution. The normal distribution has kurtosis of zero. Kurtosis characterizes the shape of a distribution - that is, its value does not depend on an arbitrary change of the scale and location of the distribution. For example, kurtosis of a sample (or population) of temperature values in Fahrenheit will not change if you transform the values to Celsius (the mean and the variance will, however, change).
- **Sample S Square (S²)** is a measure on a random sample that is used to estimate the **variance** of the population from which the sample is drawn. Numerically, it is the sum of the squared deviations around the mean of a random sample divided by the sample size minus one. Regardless of the size of the population, and regardless of the size of the random sample, it can be algebraically shown that if we repeatedly took random samples of the same size from the same population and calculated the variance estimate on each sample, these values would cluster around the exact value of the population variance. In short, the statistic s squared is an unbiased estimate of the variance of the population from which a sample is drawn.

These definitions were adapted from:

Krathwohl, D. R. (1997). Methods of educational and social science research: An integrated approach. White Plains, NY: Longman Publishing Group.

The Internet glossary of statistical terms available online at: http://www.animatedsoftware.com/statglos/statglos.htm

The Statistics.com Glossary available at online at: http://www.statistics.com/content/glossary/

Evaluation Citations

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (1995). *Standards for Educational and Psychological Testing*. Washington, DC: Authors.

Belson, W. (1981). *The design and understanding of survey questions.* Aldershot, England: Gower.

Biklen, S., & Bogdan, R. (1998). *Qualitative Research for Education: An Introduction to Theory and Methods.* Boston: Allyn & Bacon.

Campbell, Donald T., & Stanley, Julian C. (1966). *Experimental and quasi-experimental designs for research.* Skokie, IL: Rand McNally.

Charles, C., & Mertler, C. (2002). *Introduction to Educational Research*. Boston: Allyn & Bacon.

Cook, T., & Reichardt (Eds.). (1979). *Qualitative and quantitative methods in evaluation research.* Beverly Hills, CA: Sage.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Gall, J., Gall, M., & Borg, W.(1999). *Applying Educational Research: A Practical Guide*. Boston: Allyn & Bacon.

Guba, E., & Lincoln, Y.S. (1981). *Effective evaluations: Improving the usefulness of evaluation results through responsive and naturalistic approaches.* San Francisco: Josey-Bass.

Guttentag, M., & Struening, E. (1975). *Handbook of evaluation research*. Beverly Hills, CA: Sage.

Haller, E. & Kleine, P. (2001). *Using Educational Research: A School Administrator's Guide*. Boston: Allyn & Bacon.

Herman, J., & Winters, L. (1992). *Tracking your school's success: A guide to sensible evaluations.* Newbury Park, CA: Corwin Press.

Johnson, B. & Christensen, L. (2000). *Educational Research: Qualitative and Quantitative Approaches.* Boston: Allyn & Bacon.

Kazdin, A.(1982). *Single-case research designs*. New York: Oxford University Press.

King, J., Morris, L., & Fitz-Gibbon, C. (1987). *How to assess program implementation.* Newbury park, CA: Sage.

Krathwohl, D. (1998). *Methods of Educational and Social Science Research: An Integrated Approach*. Boston: Allyn & Bacon.

Merriam, S. (1988). *Case-study research in education: A qualitative approach*. San Francisco: Jossey-Bass.

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: A sourcebook of new methods.* Beverly Hills, CA: Sage.

Patton, M. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Program Evaluation Kit. (1987). Newbury Park, CA: Sage.

Rog, D. (1997). *Progress and future directions in evaluation: Perspectives on theory, practice and methods.*

Rossi, P., & Freeman, H. (1991). *Evaluation: A systematic approach*. Newbury Park, CA: Sage.

Rossi, P., Wright, J., & Anderson, A. (Eds.). (1983). *Handbook of survey research*. New York: Academic Press.

Schlalock, R. (2001). *Outcome-based evaluation*. New York: Kluwer Academic/Plenum Publishers.

Schmuck, R. (1998). *Practical Action Research for Change.* Boston: Allyn & Bacon.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. San Francisco: Jossey-Bass.

Smith, M., & Glass, G. (1987). *Research and Evaluation in Education and the Social Sciences.* Boston: Allyn & Bacon.

Sudman, S. (Ed.). (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research.* San Francisco, CA: Jossey-Bass Publishers.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco, CA: Jossey-Bass Publishers.

Suter, W. (1998). *Primer of Educational Research*. Boston: Allyn & Bacon.

Yap, K., Aldersebaes, I., Railsback, J., Shaughnessy, J., & Speth, T. (2000). *Evaluating Whole-School Reform Efforts: A Guide for District and School Staff* (2nd ed.). Portland, Oregon: NWREL.

Yin, R. (1984). *Case study research: Design and methods*. Beverly Hills, CA: Sage.

Statistical and Educational Measurement Resources

Haslam, S., & McGarty, C. (1998). *Doing Psychology: An introduction to research methodology and statistics.* Thousand Oaks, CA: Sage Publications.

Boruch, R. (1997). *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage Publications.

Loewenthal, K. (1996). *An introduction to psychological tests and scales.* London: UCL Press Limited.

Urdan, T. (2001). *Statistics in plain English*. Mahwah, NJ: Lawrence, Erlbaum Associates, Inc., Publishers.

Websites

There are a number of websites that contain information relevant to conducting evaluations.

Buros Institute of Mental Measurements http://www.unl.edu/buros/
The Buros Institute publishes a annual yearbook that reviews thousands of psychological and educational tests and measures. Their website also offer reviews and a test locator.

ERIC Clearinghouse on Assessment and Evaluation (with Test Locator)
http://www.ericae.net/
This website allows you to search for reports, books, and other documents in the U.S. Department of Education's main Educational Resources Information Center (ERIC) database, in addition to providing help in locating tests and measures.

Northwest Regional Education Laboratory (NWREL) http://www.nwrel.org/eval/index.html
NWREL offers a list of publications on research and educational evaluation online.

U.S. Department of Education http://www.ed.gov
This is the main website for the U.S. Department of Education, which offers links to state agencies, technical assistance providers, regional education laboratories, and other resources.