

# **Evaluating Standards-Based Professional Development for Teachers: A Handbook for Practitioners**

Michael Puma  
Jacqueline Raphael  
**The Urban Institute**

2001

***Prepared for:***

U.S. Department of Education  
Planning and Evaluation Service  
Collette Roney, Project Monitor  
400 Maryland Avenue, SW  
Washington, DC 20202

***Prepared by:***

The Urban Institute  
Education Policy Center  
2100 M Street, NW  
Suite 500  
Washington, DC 20037

## Contents

<b><i>Introduction. Why Use This Handbook?</i></b>	4
Some Guiding Principles Used in Writing This Handbook	5
Overview of This Handbook	7
<b><i>Chapter 1. Professional Development and Systemic Standards-Based Reform</i></b>	9
Components of Standards-Based Reform	9
Linking Professional Development to Student Outcomes	11
What We Know about the Effects of Professional Development	11
Range of Professional Development Approaches	14
<b><i>Chapter 2. Evaluation: Basic Definitions and Steps</i></b>	16
What Is Evaluation?	16
What Are the Benefits of Evaluation?	16
What Are the Types of Evaluation?	18
What Are the Steps in an Evaluation?	20
How Can You Address Concerns about Evaluation?	21
<b><i>Chapter 3. The Evaluation Frame: Understanding the Journey</i></b>	23
Gathering Information about the Program	23
Exhibit 1: Information To Collect about the Program	25
Exhibit 2: Linking Intermediate and Final Outcomes	27
Understanding the Program: The Ringwood School District Experience	28
Exhibit 3: The Ringwood School District’s Logic Model	30
<b><i>Chapter 4. Stakeholders: Deciding Who Will Come on the Trip</i></b>	31
Creating an Evaluation Team	31
Should You Use an Outside Evaluator?	32
Involve Key Stakeholders in the Evaluation	33
Know Your Constraints	34
Creating an Evaluation Team in the Ringwood School District	34
<b><i>Chapter 5. Evaluation Goals: Determining Your Destination</i></b>	36
Establishing Evaluation Goals and Objectives	36
Examples of Questions for an Evaluation of Teacher Professional Development	37
Deciding on the Destination: The Ringwood School District Experience	39
<b><i>Chapter 6. Evaluation Design: Plotting Your Course</i></b>	43
Choosing a Design Strategy — Things Evaluators Worry About	43
Using Real-World Examples To Understand Evaluation Design	46
Formative or Process Evaluation Design	46
Applying Formative Evaluation to Teacher Professional Development: The Ringwood Experience Continued	50
Types of Summative or Impact Evaluation	52

Exhibit 4: Types of Impact Evaluation Designs	53
Randomized Experiments of Professional Development	53
Applying Randomized Experiments to Teacher Professional Development: Two Examples	54
Quasi-Experimental Impact Studies	56
Applying Quasi-Experimental Designs to Teacher Professional Development: The Kramer School District Experience	57
Non-Experimental Impact Studies	60
Choosing an Impact Evaluation Design	60
<b><i>Chapter 7. Data Collection: Getting Started</i></b>	<b>62</b>
Deciding What Information to Collect	62
Exhibit 5: Linking Research Questions, Measures, and Data Sources	63
To Sample or Not to Sample: That Is the Question	64
Making the Choices: Kramer School District Revisited	69
Exhibit 6: Example of Kramer School District and Teacher Sample with Multiple Treatments	71
<b><i>Chapter 8. Data Collection: Choosing the Methods</i></b>	<b>72</b>
Deciding How to Collect the Data	72
Exhibit 7: Methods of Data Collection	74
Consider a Range of Tools	74
Data Collection Methods for Kramer School District	80
<b><i>Chapter 9. Data Collection: Creating and Using the Tools</i></b>	<b>81</b>
Constructing Data Collection Instruments	81
Pre-Testing Your Data Collection Instruments	87
Managing Your Data Collection	88
Logistical Planning for Data Collection	89
A Note about Using Data from Existing Records	91
Protecting Human Subjects and Maintaining Confidentiality	91
Data Collection in Kramer School District	92
<b><i>Chapter 10. Data Analysis: Understanding What Happened</i></b>	<b>95</b>
Getting Your Data Ready for Analysis	95
Types of Analysis	98
Getting the Story Down: Kramer School District Continued	99
<b><i>Chapter 11. The Evaluation Report: Telling the Story</i></b>	<b>103</b>
Interpretation of Results	103
Reporting	103
Exhibit 8: Sections of the Final Report	106
Where Do You Go from Here?	109
Final Thoughts	110

<i>Appendix: Guiding Principles for Evaluators</i>	112
<i>References and Resources</i>	116
<i>Glossary of Common Evaluation Terms</i>	123

## Introduction. Why Use This Handbook?

In the era of systemic standards-based reform based on challenging new academic standards, school districts, schools, teachers, and students are being held accountable for improved educational outcomes. To help meet these new goals, school administrators have an important responsibility to ensure that teachers possess the skills they need to meet higher expectations. As a consequence, many districts are changing the way they plan and deliver professional development. No longer is professional development seen as simply having teachers complete a specific number of hours of training. Instead, districts now are concerned about outcomes—typically, changes in teachers' knowledge, skills, attitudes, and/or behaviors, and, ultimately, improved student learning, as measured on tests aligned with state or district standards.

The way professional development is delivered to staff is also changing. Instead of offering only short-term in-service workshops, many districts are now providing intensive professional learning opportunities for teachers on a continual basis, such as creating mentoring relationships and building learning communities within schools. Providing these types of opportunities can be costly, however. Consequently, district staff must make informed decisions about the types of training opportunities to offer, for how long, and for which staff. This, in turn, means that districts will need to collect ongoing evidence about the effectiveness of their professional development programs and be able to revise their staff development programs to achieve the intended objectives and meet changing demands.

This handbook is written to help district staff members gain a working knowledge of how to evaluate their professional development programs. A wide range of staff may find themselves conducting or overseeing such evaluations, including directors of research and evaluation, professional development, and federal programs, as well as district staff members in small districts who must wear several hats. All of these individuals can benefit from this handbook.

Throughout, we have assumed that the district is engaged in systemic standards-based reform; that is, that it has already adopted learning standards and is using student assessments that are well aligned with those standards. We also assume that the district's professional development efforts are linked to its overall plan for focusing all components of the educational system on helping all students meet higher academic standards. This

professional development is integrated with all aspects of the district's improvement efforts.

In writing this handbook, we have also assumed that the reader is *not* trained in evaluation, and that he or she needs basic practical information about how to design and conduct evaluations. Whether the district conducts its evaluation in-house or works with an outside evaluator, we assume that the objective is to get as much out of the evaluation *process* as possible—meaning that the district will use the results to better understand and improve its professional development program.

### Some Guiding Principles Used in Writing This Handbook

Several principles guided us in the preparation of this handbook.

**Evaluating professional development programs will strengthen a district's entire reform effort.**

Many districts, overburdened with administrative and management tasks and short on funds, fail to incorporate evaluation activities into their ongoing program operations. In these situations, staff are hard-pressed to attribute observed improvements in teaching and learning to specific professional development efforts. Although some professional development programs may appear more successful than others, these districts lack a mechanism to link their “hunches” to documented results, other than anecdotal information. The absence of an evaluation can also weaken staff support for a professional development program. Participants may not receive feedback from district leaders on how professional development contributes to positive outcomes and may not understand how decisions about professional development are being made.

Evaluation results can be used to make thoughtful, cost-saving decisions about how to meet a district's professional development needs. Not only does evaluation help answer questions about the particular professional development program under study, but it builds a district's internal capacity for critical thinking, data collection and analysis, and overall decisionmaking related to reform (Taylor-Powell et al., 1998). Evaluation also helps district and school staff and other key stakeholders agree on a clear focus for districtwide reform efforts. For these reasons, we believe evaluation should be viewed as part of the process of building local capacity for reform, rather than merely as an “add-on.”

**There is no one best way to evaluate a professional development program.**

We cannot prescribe a “cut and dry” model for evaluating professional development efforts, for two reasons. First, there is no one “best” professional development program. The literature on professional development suggests that high-quality professional development programs consist of a combination of strategies selected by districts based on individual district goals and contextual factors (Loucks-Horsley et al., 1998). The process of carefully planning and conducting such a program and reflecting on the outcomes appears more important to the quality of a professional development program than the use of a particular strategy or model.

Second, there is no one “best” evaluation plan. Evaluations need to be sensitive to local programs, and therefore no simple recipe exists for how they ought to be conducted. Those responsible for the evaluation must make a series of choices based on key features of the program, including its purposes and objectives and local context, stakeholders’ needs, available resources for the evaluation, and practical constraints on data collection. Our goal with this handbook is to guide you in making these decisions.

**To conduct a successful investigation, evaluators of professional development programs will need to balance conflicting demands from stakeholders**

Tension among the needs of different stakeholders may emerge during evaluations of professional development programs. Some stakeholders may focus on getting to the “bottom line,” seeking to isolate the effects of a specific program on student test scores. Others may be far more interested in understanding how and why a program is (or is not) working and the program’s relationship to meeting the goals of broader systemic reform efforts. All of these needs are important, but with limited evaluation resources, evaluators may struggle to strike a balance.

Another tension inherent in professional development evaluation is between “intermediate” and “end” outcomes. Professional development benefits teachers, students, and in some cases the school more generally. But if the bottom line, ultimately, is the impact of a program on student achievement, the question becomes how much weight to give other intermediate effects, such as changes in teachers’ knowledge, attitudes, and behavior. Should changes in instruction be considered an intermediate or an end outcome? The

answers to these questions depend on the goals of the professional development program and the evaluation, as well as on constraints on the evaluation's scope.

Finally, stakeholders may demand evaluation results in a short time frame, such as one year or less. However, it is likely that the deeper changes anticipated for professional development that is related to new, higher academic standards—particularly the end outcomes related to students—will take longer to achieve. During the first year of a professional development effort, for example, changes in teacher attitudes and behavior may be a reasonable outcome. On the other hand, a year or two later may be an appropriate time to begin investigating changes in student achievement. Evaluators must, therefore, balance the short-term need for results with the overall focus on developing a systemic reform effort, expecting to measure important overall goals over a longer period of time.

## Overview of This Handbook

This handbook walks the reader through the decisionmaking that is involved in the evaluation of professional development. Although district staff can use this handbook throughout the evaluation process, we recommend reading it in its entirety before beginning to design and implement an evaluation project, as many issues covered later in the book are relevant to the planning and design phase. Most likely, readers will also want to consult other resources about evaluation; some suggested resources are listed in the References section at the end of the handbook.

Chapter 1 provides a brief overview of the role of professional development in systemic standards-based reform, and chapter 2 provides general definitions and a description of the steps needed to plan and conduct a professional development evaluation. Starting with chapter 3, we move through the conceptual issues underlying evaluation, including designing the evaluation, collecting and analyzing data, and reporting results. To help illuminate key steps in the process, we use fictitious school—but “real life”—examples of districts throughout the handbook. Two of the examples are extended: The Ringwood School District's experience establishing evaluation plans is discussed in chapters 3 through 6, and the Kramer School District's work in designing and conducting an evaluation is discussed in chapters 6 through 10. Finally, appendices at the end of the handbook suggest printed and electronic resources that may be useful to you in planning and conducting evaluations of professional development programs, and the glossary defines terms that are commonly used to discuss evaluation.

Using this handbook will help users understand the key steps in the design and implementation of a district evaluation of professional development. Throughout this handbook, we will compare this process to that of planning and going on a school field trip. We describe the following key steps:

- Understand the Journey: Describe the program to be evaluated**
- Decide Who Will Come on the Trip: Identify stakeholders and involve them in the evaluation**
- Determine Your Destination: Establish your evaluation goals and objectives**
- Plot Your Course: Write your evaluation design**
- Gather Information: Identify your data sources**
- Gather Information: Choose your data collection methods**
- Gather Information: Create and use data collection instruments**
- Understand What Happened: Analyze your data**
- Tell the Story: Interpret and report your results**

## Chapter 1. Professional Development and Systemic Standards-Based Reform

When a district undertakes standards-based reform, it is changing the whole education system. This is a lot of work. For example, the district has probably established content and performance standards, disseminated a new curriculum to teachers and principals, and begun to plan for or administer new state tests as part of a new accountability system. These changes pose significant challenges for teachers. When they voice their concerns, you—as a district or school administrator—point out that the district is offering many new professional development opportunities. But the teachers cry, “When do we have time?” Although you understand the frustration of teachers and other school staff, at the same time you ask, “What’s so new? Haven’t we always been trying to improve instruction?”

What makes *systemic* reform so challenging—for teachers, students, school staff, and the community served by the school district—is that this type of educational reform demands significant changes to several components of the educational system (e.g., expectations, curriculum, instruction, assessment) that are often expected to occur simultaneously. To be systemic, these changes to multiple components of the educational system must be coordinated; that is, they must be driven by a common goal of improved learning for *all* students, and there must be linkages among the various parts. For example, if teachers of early elementary grades do not articulate a smooth, developmentally appropriate sequence of learning, then students moving through the school can get caught in a literal “learning gap.” (The same can be said of the need for between-school collaboration within a district.) This work can be difficult.

### Components of Standards-Based Reform

On the surface, the concept underlying standards-based reform seems fairly straightforward. Yet it has been developed over a long time period. Starting soon after the National Commission on Excellence in Education’s 1983 publication of *A Nation at Risk*, which described a “rising tide of mediocrity” in U.S. schools, major efforts to improve education were launched by states and throughout the nation. State efforts culminated in the 1989 “education summit,” which subsequently led to passage of two pieces of federal legislation to stimulate school reform. The first, the Goals 2000: Educate America Act, passed in 1994, provided funds to states and districts to support systemic reform efforts

based on state standards. In addition, the reauthorized Elementary and Secondary Education Act of 1994 (the Improving America's Schools Act) made standards and accountability an integral part of all of the federal programs.

The main shift from previous federal legislation was a greater focus on state leadership as the driver of school reform, and the need for aligned and coherent policies regarding standards for what students are expected to learn, the content of instructional materials and curriculum, classroom pedagogy, teacher preparation, and accountability and assessment systems. The expectation was that (1) states would establish challenging content and performance standards for all students, and (2) states and districts would, in turn, align other parts of the education system with these standards; that is, that student assessments and accountability systems would be developed that are aligned with the articulated standards, teachers would be provided with the necessary training and support to help them achieve the standards, and schools and students would be held accountable for attaining the desired level of proficiency. This "systemic" alignment was expected to focus school improvement efforts on improved teaching and learning, and yield academic gains.

What made this idea so radical was that it sought to fundamentally alter the way educational changes were traditionally made by moving from an incremental approach—adjusting a single component of the instructional process such as reducing class size—to a systemic perspective, in which reforms change how the different components of an educational system work together. The underlying hypothesis was that increasing the alignment or coherence among the different components, actors, and agencies that make up the education system would make schools more effective, and this, in turn, would provide the impetus needed to drive the overall system toward higher levels of student learning.

As a consequence, professional development for teachers in districts engaged in systemic, standards-based reform is often organized around efforts to address multiple components of the education system. This is what is meant by an *aligned* professional development program: it is integrated with other changes taking place in the district that are intended to work together to improve teaching and learning through the implementation of standards.

## Linking Professional Development to Student Outcomes

In districts engaged in systemic standards-based reform, the ultimate goal of professional development is to contribute to increased student learning, as typically measured by state assessments that are aligned with district or state academic standards. The extent of alignment between the professional development and other components of the reform effort—including standards, curriculum, assessments, and accountability—is critical to the contribution of the professional development to systemic reform. If the professional development program is *not* well linked to the goals of the overall system, then teachers may well become confused about the district's expectations for them, and this confusion can result in less productivity and/or no change (or even negative change) in student learning. In addition, systemic reform requires that necessary changes be implemented on a large scale. Thus, districts must eventually deliver high-quality professional development related to standards to most, if not all, of its classroom teachers.

An important “link” in the chain connecting professional development and student learning outcomes is the effect of professional development on its direct participants—typically teachers, although other staff (e.g., teaching aides, administrators) may be included. This is a necessary step. Many district evaluations of professional development limit the scope of their evaluations to the effect of professional development on teachers, without going on to assess the effects on students taught by participating teachers.

Other forces affect the relationship between professional development and student learning, such as school policies, school administrators, and parents. These factors can have significant effects on student learning—and can be instrumental in facilitating or inhibiting improvements in teaching and learning. To evaluate the professional development program's success in raising student achievement, you may need to learn how these factors played out during the period of the evaluation.

## What We Know about the Effects of Professional Development

According to most education researchers, we have very little hard evidence about the effects of professional development on teaching and learning (Birman et al., 1998; Donnelly et al., 2000; Elmore and Rothman, 1999; Wilson and Berne, 1999). However, some research does suggest that intensive and long-term professional development may be more effective at changing teacher practice than traditional forms of professional development (Birman et al., 1999; Cohen and Hill, 1998; Shields, Marsh, and Adelman, 1998). For example, a study conducted by the Consortium for Policy Research in

Education (Corcoran et al., 1998) for the National Science Foundation suggests that a minimum of 100 hours of contact time is needed for a professional development program to have its intended effect on instruction.

Cohen and Hill (1998), in a study of mathematics reforms in California, found that professional development that focuses on the specific curriculum teachers will use in the classroom is more effective at improving both instruction and student learning than more general professional development. When such curriculum-based professional development also emphasizes other instructional elements of the system, such as student assessment, Cohen and Hill found it to be even more influential.

Similarly, a review of professional development in mathematics and science conducted by Kennedy (1998) suggests that staff development that emphasizes how students learn specific mathematics and science concepts is more effective in increasing student achievement than professional development focused on general teaching principles. Kennedy also found that the *content* of professional development opportunities in mathematics was the most important variable related to program effectiveness, even when compared to popular structural features of “new” professional development approaches such as extended contact time, in-class visits from coaches, and whole-school staff development.

Despite the need for more research on the effects of professional development, educators are calling for the replacement of the standard in-service workshop with more promising professional development opportunities. In response, professional and research organizations have developed standards, or key principles, to guide the design and implementation of professional development programs (NCTM, 1989, 1991; National Council of Teachers of English, 1996; National Board for Professional Teaching Standards, 1989). These recommendations are often based on commonsense notions about what type of professional development is likely to provide teachers with in-depth learning opportunities.

Below, we present a compilation of features of high-quality professional development identified in several prominent research sources.<sup>1</sup>

---

<sup>1</sup> Sources include a literature review conducted by SRI International (Donnelly et al., 2000) for a project for the U.S. Department of Education on professional development in educational technology. The researchers combined the findings from a previous review of professional development conducted by Corcoran, Shields, and Zucker (1998) on the National Science Foundation’s State Systemic Initiatives with their own review of over 12 additional articles and reports by research organizations and researchers. In addition, the work of Birmanet al. (1998) on the

### **High-Quality Professional Development:**

- **Promotes an approach to teaching and learning that supports high standards for all students.** These approaches are aligned with standards and assessments. They can incorporate strategies for meeting the educational needs of diverse student populations. These strategies must be grounded in established knowledge about effective classroom teaching and learning and must be accessible to all educators.
- **Increases teachers' knowledge of specific content and of how students learn that content** Deepening teachers' knowledge of specific disciplines that they teach is critical. Also important is the development of "pedagogical content knowledge"—professional development that focuses on the pedagogical implications of the discipline, such as understanding how students learn the discipline at different ages and in different contexts. Such professional development is rigorous and based on the knowledge base about teaching, as well as the underlying theory for that knowledge base.
- **Provides intensive, continuous, in-depth learning opportunities for teachers, with follow-up and support.** Professional development should include a high number of contact hours and span a long time period. These experiences should build on existing knowledge and permit teachers to collaborate, learn from each other and from external sources, experiment with new techniques, gain critical feedback, and continue to refine their teaching processes over a significant time period, in a continuous fashion, with repeated follow-up and support for ongoing learning as needed.
- **Expands the traditional role of teacher.** Current reforms demand that teachers take on new responsibilities to become leaders, mentors, peer coaches, curriculum and/or assessment designers, planners, and facilitators. In this environment of reform, teachers and other instructional staff form a community of learners who plan and work together to solve problems across the school and/or district. In addition, as many districts devolve authority to the school level, teachers are being asked to assume new roles in school governance and management (Corcoran, 1995). Teachers may be involved in identifying their professional development needs and in planning, designing, and delivering opportunities to meet those needs, as well as in assessing the effectiveness of these opportunities.
- **Connects directly to other reform programs and initiatives.** Professional development in the context of standards-based reform must be linked to other federal, state, district, and/or school initiatives. Such linkages can help to support teachers implementing new practices. The connection to school reform is also important to guarantee that professional development reflects specific local needs and abilities.
- **Is accountable for results.** Professional development should be evaluated regularly for its effects on teaching and learning. Multiple sources of data (e.g., teacher portfolios, classroom observations, peer evaluations, student performance) should be used, with data collected at different times during the program implementation process. The results of these evaluations should be used to support continuous improvement.
- **Is collaborative.** By working together, teachers break down the isolation of individual classrooms and can begin to transform a whole school. Professional development activities should occur in groups of teachers from the same school, department, or grade level.
- **Is active and focused on problem-solving.** Teachers need to be actively engaged in teaching and learning, particularly through curriculum development, action research, and other problem-solving activities.

---

evaluation of the Eisenhower Professional Development Program; the National Partnership for Excellence and Accountability in Teaching (1999); and Thomas Corcoran (1995) were used to prepare this list.

In addition, two key components link professional development to standards-based reform and make it systemic:

- (1) ***Enhancing teachers' knowledge of the subject matter referenced in the standards and their knowledge of how to teach that content to students***  
This can be accomplished most obviously by ensuring that the content and approach to teaching subject matter is well connected to the standards. Other approaches involve a greater connection between professional development and the other instructional elements of the education system, such as professional development that involves teachers in producing and learning to use aligned curricula and/or assessments more effectively.<sup>1</sup>
- (2) ***Supporting teachers in acquiring and using their new knowledge by making necessary organizational and administrative changes*** (e.g., providing additional resources and/or additional time, greater collaboration in the school and/or district, enhanced roles for teachers). Such changes, as mentioned, are critical to the systemic view of the school as part of a larger system that influences classroom instruction.

### Range of Professional Development Approaches

Districts engaged in systemic standards-based reform use a variety of approaches for the delivery of professional development, including the following:

- ***Coaching and mentoring relationships between teachers or professional developers.*** Typically, these relationships involve working one-on-one with an equally or more experienced educator on issues related to pedagogy such as classroom observation and feedback, collaborative planning, and team teaching.
- ***Train-the-trainers approach.*** This approach encompasses the development of the knowledge and skills of selected staff members who go on to train other teachers, usually at their schools, by presenting workshops, doing demonstrations, and supporting teachers' growth in other ways. This approach can be more cost-effective than providing professional development to an entire staff.
- ***Collaborative study groups.*** In these groups, teachers meet with staff from outside the district in professional networks, or with staff from within the district in study groups (e.g., grade-level or subject-area groups). Such collaborative forums allow teachers to explore and discuss topics of common interest and share information, strategies, and long-range plans.
- ***Learning communities.*** In learning communities, groups of teachers and staff at a school focus on self-assessment and the analysis of teaching/learning data and provide many opportunities for professional growth. In districts that promote school-based management, principals and school staff work together to develop their own learning communities, including identification of school-based professional development goals and activities.

- **Learning in professional development schools.** These schools, established by university/school partnerships, explore and model research-based teaching techniques. Pre-service and in-service teachers are served in these laboratory-like settings, frequently in individualized programs.
- **Reflection on student work samples.** In this practice, participants examine student work to better understand how students learn and how well they have progressed in achieving the standards. Teachers and other staff also refine their learning expectations for students by identifying what constitutes proficiency as measured against the standards.
- **Participation in curriculum development and on other curriculum-related tasks** These tasks include teacher development of curricula and/or instructional materials aligned with the standards, as well as focused work on curriculum implementation and review and evaluation of instructional materials. Such activities develop teachers' capacity to make the standards operational.
- **Action research.** Participants conduct research based on what happens in their classrooms. Action research gives staff an opportunity to systematically explore questions directly related to their own needs, and to tailor the learning experience considerably. Action research has the potential to renew teachers' commitment to professional growth.
- **Workshops, courses, and seminars.** The traditional format for professional development is a structured setting that provides an opportunity for concentrated learning, typically delivered by experts.
- **Institutes** These intensive professional development opportunities are often held for one or more weeks during the summer. Institutes typically involve hands-on work that gives teachers practice in how to approach instruction differently.

Rather than conceptualizing professional development as a self-contained effort to improve instruction, many districts engaged in systemic reform—like individual schools—strive to function as *learning communities*. These districts define their long- and short-term goals based on their standards, plan and implement programs to achieve those goals, and then assess their results using assessments aligned with their standards. Such districts plan professional development based on needs identified in this process. Such a *process* for designing professional development represents an opportunity to integrate evaluation into a continuous improvement model for the district.

## Chapter 2. Evaluation: Basic Definitions and Steps

### What Is Evaluation?

***Evaluation is the “systematic assessment of the worth or merit of some object” (Scriven, 1967).***

## S

Scriven's definition of evaluation conveys the underlying idea that, despite the variety of types of evaluation, at some level all evaluations are intended to make judgments about the “object” being evaluated—typically a **program**, by which we mean a set of activities (e.g., components of a professional development program), supported by a variety of inputs or resources (e.g., staff, equipment, money), that is intended to achieve specific outcomes (e.g., teaching skills aligned to standards) among particular target groups (e.g., classroom teachers). Combining the concept of evaluation and the definition of a program, Cronbach et al. (1980) define **program evaluation** as the “systematic examination of events occurring in ... a contemporary program ... to assist in improving this program and other programs having the same general purpose.” The key points to be kept in mind, then, are (1) evaluation is a *systematic* endeavor that (2) involves *acquiring and assessing information* to (3) *influence decisionmaking*. In other words, evaluation is about providing **data** that can be used to make a decision, to establish a new policy, or to take a specific goal-directed action.

### What Are the Benefits of Evaluation?

Evaluations take time and resources, so why should you want to evaluate your professional development program? You may even think you already know it works! However, even when a program appears to be effective, the information you acquire through evaluation helps you and others gain a better understanding of your program's effect on your teachers and, ultimately, on your students. This information can, in turn, help you improve your training and make it more efficient. In addition, evaluations can provide information to a variety of people and organizations that are interested in what you are doing, including sponsors and/or donors (state officials, your school board, the district superintendent, and external funders), target groups (teachers and other staff), administrators, and other individuals with a stake in the results of your program (e.g.,

parents, students, and the community). If used properly, evaluations can lead to increased success for managers and staff and can result in service improvements for participants. Evaluations can serve many different purposes. They can help program managers and staff determine *what services they need to offer, how well they are providing these services, and the likely consequences of their efforts*. Which of these questions you will answer depends on the specified goals and objectives of your evaluation. How well they are answered will be determined by the quality of the research strategy, how well it is implemented, and the level of resources available to support the evaluation.

**Evaluations can be used to answer many different types of questions, including issues of *program merit* (i.e., what is the quality of the professional development program? can we improve it?), of *worth* (i.e., is the staff development program cost-effective? can the same or better results be achieved at lower cost?), or of *significance* (i.e., how important are the effects of the program? does the professional development make a difference for teachers? for students?).**

Evaluations come in different forms, and the information they produce can serve different purposes. Here are some examples:

- ***Identifying service needs.*** Evaluations can provide data on the professional development needs of your staff (i.e., determining the knowledge and skills your teachers have learned and need to learn), and can help you decide which can be addressed by existing services and which will require the creation of new initiatives. For example, your district may want to conduct a series of evaluation activities to determine what skills your instructional staff need to acquire to develop curricular units aligned with subject-matter standards.
- ***Trying out a new program.*** In some cases, districts may want to experiment with a new or innovative approach before deciding whether to implement it districtwide. This is like a clinical drug trial that is used to determine if the new method is better (or worse) than existing therapies. For example, your district may want to compare the use of summer workshops, training classes during the school year, and an ongoing mentoring model to determine which is the best vehicle for districtwide implementation. These options can be evaluated on a “pilot test” basis, using a relatively small number of teachers. The results of this first-stage evaluation would then lead to the selection of a strategy for broader implementation that can subsequently be evaluated as it is rolled out to all your teaching staff.
- ***Tracking program implementation and interim accomplishments.*** Once a professional development program has been implemented, evaluations can be used to keep track of program activities. These types of data, often used by administrators and found in “management information systems,” help managers to monitor progress against goals and to adjust programs as needed to improve their effectiveness. The types of information you might collect for this purpose could include the number of teachers that have received professional development, and the number that have acquired the desired knowledge or skills.

- **Assessing the achievement of program goals.** Beyond determining if particular program activities are being implemented as planned, evaluations can also be used to determine whether the overall program achieved its intended purpose. For example, frequently asked questions include, “Did the professional development make a difference; that is, did teachers improve their classroom instruction, and did student achievement increase?” “Under what circumstances were the goals met, and for which participants (e.g., differences across schools, or types of teachers)?”

An evaluation can be designed to produce one or more of these different types of information. The decision about what information is desired, however, will have important implications for the design of your evaluation, as will be discussed in subsequent chapters.

## What Are the Types of Evaluations?

Although evaluation has many uses, for convenience evaluators often group evaluations into two broad categories: **formative** and **summative evaluation**.

**According to Stake (1967), “When the cook tastes the soup, that’s formative evaluation; when the guests taste the soup, that’s summative evaluation.”**

### Formative Evaluation

Formative evaluation—also known as process, or implementation, evaluation—involves monitoring the implementation or operation of a program, especially a new one. The types of questions typically asked in a formative evaluation include the following:

- **Did the program (or particular activity) occur as envisioned?** If not, what barriers or obstacles prevented it from being executed? For example, in studying professional development, we may want to know whether appropriate personnel delivered services, whether staff attended and for how long, whether they completed the program, and the quality of the training in relation to the program’s objectives.
- **To what extent were activities conducted according to the proposed timeline?** Were all the teachers trained by the expected deadline?
- **To what extent are actual program costs in line with budget expectations?** Did the training cost more than originally planned? If so, why?
- **To what extent are participants moving toward the anticipated goals of the program?** Did the staff find the training useful? Do the teachers report gaining something new? Do they plan to use it in their classrooms? What impediments to implementing what they were taught were encountered by the teachers?
- **Did the activity lead to the expected change in organizational operations?** For example, do teachers, do anything differently in their classrooms?

For many districts, answers to these process evaluation questions meet significant information needs of managers and staff. If done well, such evaluations can tell them if the program is operating as anticipated (or desired) and how to make any needed changes.

### Summative Evaluation

Over time, program managers, and often funders and program sponsors, will want to know more than what the results of formative evaluation can provide: Does an investment in professional development for teachers result in changes in student learning, particularly as reflected in improvements in student test scores? This type of question refers to what evaluators call “program impact,” and requires an ability to attribute any observed changes to the effect or “impact” of the program. By impact we mean something more than a correlation between program implementation and changes in teachers and/or students. Rather, we mean a determination of *whether the changes would have occurred in the absence of the program activity*. This is also called a “summative evaluation” and addresses the following types of questions:

- ***To what extent did the program meet its overall goals?*** Did the teachers gain the desired skills? Did the program change classroom instruction? Did it lead to improved student outcomes?
- ***Was the program equally effective for all participants?*** Did some teachers do better than others? Are results better for some subjects or grades?
- ***What components or activities were most effective?*** For example, does ongoing coaching and mentoring lead to better outcomes than formal professional development programs?
- ***Did the program have any unintended impacts?*** For example, has staff morale changed because of the focus on standards? Has instruction become too tied to the test?
- ***Is the project replicable? Can it be “scaled up”?*** If the program was implemented in only a few schools, or with only some teachers, can it be expanded districtwide?

Demonstrating that a particular program caused specific changes in outcomes is not easy, largely because a variety of factors, other than the particular intervention, can affect the outcomes of interest. For example, student achievement can rise (or fall) due to the effect of changes in the types of students coming into the school from year to year. These changes can occur whether the teachers received professional development or not. As will be discussed in subsequent chapters, the choice of an evaluation approach can have a

great deal to do with the extent to which one can rule out such competing reasons for any observed changes in program outcomes.

## What Are the Steps in an Evaluation?

Planning for, and conducting, an evaluation involves a number of steps that can seem complex but are, in many cases, conceptually rather simple. In many ways, it can be like planning a school field trip:

- **Understand the context.** Field trips are not stand-alone exercises, so any good teacher has to understand the overall course of study before deciding how to add pieces to the instructional process. In a similar fashion, evaluators need to understand the program that is being evaluated before deciding how to “add on” an evaluation component. You cannot evaluate what you do not understand.
- **Decide who will come on the trip.** Field trips can involve just a few students and a teacher, but larger excursions need a variety of people with different skills and resources—a bus driver, escorts or chaperones, and guides and docents. Similarly, very small evaluations can be designed and carried out by a single person, but larger projects require that you assemble a team with the right skills and resources to ensure that you are successful.
- **Determine your destination.** Obviously, a field trip requires a destination and some idea about what the students are expected to learn from the experience. In the same way, your evaluation has to have a specified “destination”—that is, you have to determine goals and objectives for the evaluation in advance, as this will guide all of your other decisions, and you have to have “research questions” that indicate what you expect to learn from the evaluation.
- **Plot your course.** Knowing where you want to end up is not enough. You have to know how to get there, how to make the trip as enjoyable and productive as possible, and what hazards to avoid. In evaluation, a well thought out plan is your most important asset. Without it, you cannot be assured that you will be able to answer the questions you want to answer in a way that can support subsequent decisions and actions. In some cases, midcourse corrections are needed to respond to changing conditions, and a backup plan can be helpful in case things do not go as expected.
- **Gather information along the way.** Since a field trip is about gathering information, students will typically have particular things that they are expected to learn about or investigate while on the trip. Evaluators, too, need to collect information along the way, and this information has to be inextricably linked to the research questions that one seeks to answer.
- **Understand what happened.** Good teachers do not want their students to simply collect information, but to learn from the experience in order to reach a deeper understanding of the subject matter. Similarly, evaluators need to analyze and synthesize the information they collect as part of their evaluation to better understand what's going on with the program. Data alone do not represent knowledge.
- **Tell the story.** Often, the best part of a field trip is being able to tell a good story, and this need to tell the story is also true for an evaluation. Without it, no one will know about the experience, or be able to use the information for later decisions and/or actions.

In the next chapters we take a careful look at each step. But before we move on, let us consider some of the reservations that district staff (and particularly the users of evaluation information) may have about this process. Being able to respond to these concerns will help you immeasurably during the early planning process.

## How Can You Address Concerns about Evaluation?

Despite the value of evaluations, some individuals are likely to object to the process. Some may fear that the evaluation will show that the program is not working as expected or that it is not being implemented well. In addition, evaluations require the use of scarce resources (money that staff may feel is better spent on services) and challenge program staff who may not know how to carry out an evaluation or use its results. To help reduce such concerns, it is advisable to do the following:

- ***Involve staff members in the evaluation process.*** Participation in the process helps keep staff from feeling that their individual performance is being scrutinized or that either their job or the entire program hangs in the balance. Staff members who participate in the evaluation are also less likely to believe that the methods used for the evaluation do not adequately capture what they or the program do. Staff members can also offer important suggestions for how to make the evaluation more efficient and less burdensome.
- ***Start the evaluation planning process early.*** Careful evaluation planning can have many benefits, including reducing the burden on participants and those delivering services, containing the costs of the evaluation, and ensuring that the evaluation meets the needs of a variety of stakeholders.
- ***Emphasize to school staff the benefits of evaluation, including improving the program.*** This may help avoid the perception that the evaluation is drawing away scarce resources (including time) better used for service delivery. Evaluations do take time and money, but a strong evaluation can help to improve efficiency later and ultimately lower program costs. In many cases, results may be needed to justify continuing the program. Staff may believe that evaluations are useful to those “checking up” on their progress, but not to them. Staff need to be shown that evaluation will be a benefit in the long run, especially to program participants if services are improved.
- ***Conduct the evaluation in a credible manner, and interpret the results carefully.*** Evaluation results can be misused, especially in a heated political environment. Stakeholders can develop strong opinions about programs, becoming less open to the true meaning of evaluation results. By being thoughtful and professional in conducting the evaluation, you will encourage objectivity among your audience. In addition, presenting results clearly and unambiguously also helps to reduce the potential for misuse. It is important to stress that there are no “negative” results; rather, evaluations yield information that can help managers and staff work more effectively.

- ***Take the time to help district and school staff understand the general concept of evaluation.*** Staff and managers may avoid evaluations because they don't understand them or don't know whom to ask for help. They may believe that evaluations are too complicated to be of use. Although the technical aspects can be complex, the evaluation process often just systematizes what many managers do already: figuring out whether the programs goals are being met, seeing what is and is not working, and so on.

The important point here is not that evaluations should be avoided, but that evaluators should be aware of these concerns and build into their plans the specific actions mentioned above to reduce possible negative consequences.

It is difficult to assess a professional development program's success, as evaluating success involves tying changes in teacher practices to changes in student learning as measured by aligned tests. However, we believe that once you become familiar with the process of evaluation—the subject of the next chapter—you will see that the benefits of evaluating your professional development program far outweigh the effort required to overcome this challenge.

## Chapter 3. The Evaluation Frame: Understanding the Journey

***One of the important things to do in any evaluation is to be sure you understand the program that you are evaluating. You cannot begin to design an evaluation for a program you do not understand!***

Teachers often use field trips as a way to enrich their classroom instruction. But to be effective, such real world experiences must be tied to the overall curriculum the teachers are using in their classrooms. Similarly, to do a good evaluation, you need to have a clear understanding of all of the components of your program, as well as the underlying logic that guides the district's thinking about how the professional development program can contribute to systemic standards-based reform. If you are responsible for planning or delivering professional development for schools in your district, then you probably know a good deal about the program to be evaluated. If not, you will want to engage in the discussions described in this chapter. Furthermore, all staff members who will be conducting the evaluation must be cognizant of relevant information.

### Gathering Information about the Program

Before you start planning your evaluation, you will want to know about the origins, goals, and context of the program. For example, the needs the program was designed to address, who the targeted audience is, the planned program activities, contextual information, program goals, and any factors that could influence levels of participation and program success. In addition, if baseline data—information gathered prior to program implementation—was collected (e.g., as part of a needs assessment), this information should be used to help frame the evaluation.

You probably already know some of this information. If not, most of it may be available in written materials about the program, such as program files from the professional development provider (if you are using an external trainer), written curricula, or professional development plans that specify key goals and objectives. In addition, you may want to meet with several individuals to gather a variety of perspectives about how the program is expected to function. Your district's director of professional development may know a good deal about the program's overall goals, for example, but a lead teacher or an outside consultant who may have trained the lead teachers may know the most

about how teachers are expected to move from exposure to new ideas, to changes in instruction, and finally to changes in student achievement. The amount and type of people to whom you should speak obviously depends on your particular program, as well as the resources you have available to devote to this task. This is an important step, but do not get too bogged down in small details. At this point you should be trying to get the “big picture.”

To guide you in this initial step in the evaluation process, the following is a list of some of the types of information you will want to collect about your professional development program:

## Exhibit 1: Information To Collect about the Program

- ***The needs the professional development was designed to address.*** These needs can be broad or simple. However, you will want to know why and how the program came into existence. If you are planning your evaluation as you plan your program (the ideal case), then this information will be key in your program-planning phase. In any case, these needs should be well documented and understood by evaluators so that they can be incorporated into the measurable outcomes identified for the evaluation.
- ***Target population(s).*** It is important to understand exactly who the target population is. Which teachers, and which students? Also, you will want to know what, if any, plans exist to “scale up” the program to more teachers and/or schools. Effecting change throughout a system—the key to systemic reform—may mean involving large numbers of teachers in professional development programs. And yet the timing for participation, and the expectations for different subgroups, may differ. Furthermore, by target population, we mean not only the primary clients (typically teachers and instructional aides), but also who was expected to be affected by the outcome of the program—in most cases, the students. You should learn about the targeted population, asking questions such as the following:
  1. Were *all* teachers targeted? Including special needs teachers (e.g., special education, English as a Second Language) and teachers of nonacademic classes (e.g., art, music, physical education)?
  2. Were other staff included (e.g., building and district administrators, counselors, librarians)?
  3. Was participation staggered, so that some participants started earlier than others? How were these decisions made and implemented?
  4. Were specific groups of participants targeted? For example, were teachers of low-performing students the “real” targets, even if all teachers were trained? Were new teachers expected to participate?
- ***Specific program activities.*** This is the easy step: listing the professional development activities. These can include provision of research experiences for teachers, curriculum development workshops, coaching in a new reading program, or technology training. It is most helpful to describe your professional development program activities in measurable terms, including the number and type of participant for each activity, and to describe whether program activities occur separately or simultaneously.
- ***The time frame.*** You will want to know the timing for program implementation and expected achievement of outcomes. For example, when did the program truly start (not the “proposed” start date)? Did different groups of participants have varying startup times? Did all participants receive the same amount of training? Furthermore, you will need to know when the anticipated intermediate and final outcomes were expected to occur. This is where your “informants” can be helpful, as they may know how much time it really takes for teachers, for example, to master the new skills covered in the professional development, and, perhaps, when the district could reasonably expect to see students benefiting from the professional development (as well as in what tests, subjects, and grade levels).

- **Factors that could influence program participation and program success.** Various aspects of the state or district policy context may be relevant to your understanding of the program. For example, if the state suddenly decides not to develop new statewide student assessments as scheduled, practitioners who have been preparing for the change may become discouraged and unenthusiastic about new programs, including professional development. Features of your student population—particularly in a given year—may be relevant, as when a large number of limited-English-proficient students enter the school district one year.
- **The costs.** The cost of the program includes time, the use of existing resources, the purchase of new resources, and changes in staff roles and responsibilities. To the extent possible, you should quantify costs for each stage of the program, including planning.
- **Expected participant outcomes.** What are the goals of your professional development effort? Typically, it is easier to think about the types of professional development activities a project will offer than the expected outcomes that are to be achieved. To identify outcomes, you may need to ask yourself and others, “What would ‘success’ as a result of this program look like?” Is increasing staff morale and motivation sufficient? Or are you expecting to see changes in the knowledge and skills of the teachers as well? And are you expecting to see these changes translated into better student outcomes?
- **Organizational changes.** Sparks (1993) reminds us that in addition to individual (teacher and student) outcomes, professional development programs may have organizational goals as well—for the district, school, or department. Such changes will affect the structure and policies of the organization, such as the reallocation of district funds or the creation of a task force to explore the use of data to drive the school improvement process. Often, these organizational changes are expected to support the achievement of participant outcomes.

In recent years, evaluators have developed a tool for organizing some of this information and, in particular, to create a picture of the ways in which program activities are linked to anticipated outcomes. This tool, called a **logic model**, describes the program to be studied and helps evaluators meet two objectives. First, it can guide decisions about how to conduct the evaluation. Second, it can be used as a tool to engage staff and participants in a dialogue about the program, providing a reality check on expected program outcomes.

In its simplest form, a logic model consists of the assumptions, or “if/then” statements, that describe the underlying logic about how your professional development activities may lead to your anticipated outcomes. For example, *if* your professional development program helps teachers create better lessons tied to state math standards, *then* your students should be learning math better, which *then* should lead to higher student test scores on your state math assessments. (See the end of this chapter for a hypothetical district’s

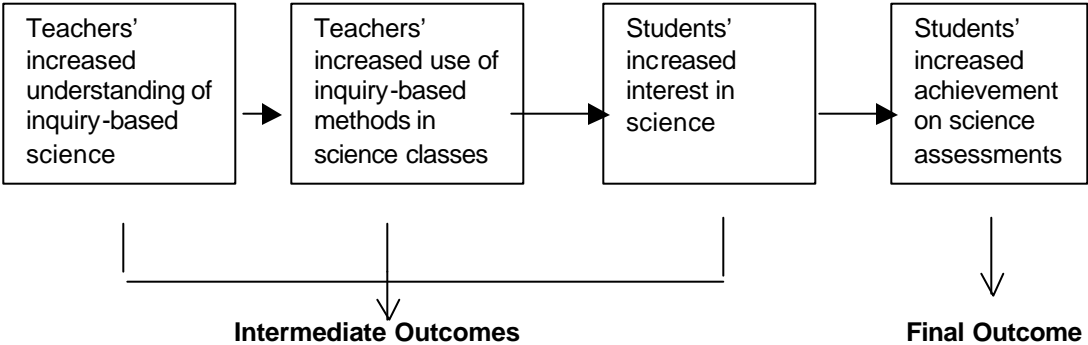
approach to developing a logic model.) In addition to helping you understand the logic of your program, creating a logic model will help you to learn about other aspects of the program, including its stage of development (“Is it a new or a mature program?”); the relevant context, including how it fits into the larger organization and broader school reform plans; planned day-to-day activities and operations; and the types and amounts of different resources that will be used. A logic model will also help you identify the sequence of **implementation** steps for your program, as well as the sequence of **outcome** steps that are expected to lead to improved student achievement.

To create a logic model for a professional development program, you will need to distinguish between “intermediate” and “final” outcomes:

- **Intermediate outcomes** are results that occur immediately or soon after the program is completed. Generally, these intermediate effects are represented by changes in participants’ (e.g., teachers and other school staff) knowledge, attitudes, and behaviors related to instruction.
- **Final outcomes** are the anticipated *consequences* of the program, or what we have referred to as *impacts*. For the purposes of professional development, this is usually changes in *student* learning or changes in student behaviors that support learning (e.g., classroom engagement).

Logic models can help you recognize that professional development efforts are designed to produce various outcomes, or effects, on both direct and indirect participants and/or organizations. For example, we typically assume that teachers’ knowledge must change for them to subsequently change their classroom practices. We then expect to see changes in student learning, as measured by aligned tests. You will want to be sure that you include in your logic model *all* of the outcomes you expect to achieve, especially those that you intend to measure in the evaluation. For example, you may wish to insert an intermediate outcome, such as increased student interest, that you believe will precede students’ increased learning. Here is an example of how this would look:

**Exhibit 3 : Linking Intermediate and Final Outcomes**



To help illustrate what the goal of the process of understanding the logic of a professional development program, here is an example of how a hypothetical district building a logic model for professional development, an example of a hypothetical district's experience follows.

### Understanding the Program: The Ringwood School District Experience

In 1998, Ringwood School District<sup>2</sup> instituted a five-year strategic plan, developed with the participation of many stakeholders, to increase student achievement 30 percent by 2002. To achieve this goal, district staff from the offices of professional development and curriculum and instruction collaborated in 1999 on the development of a professional development plan. According to the plan, by 2002 all teachers who had been in the district at least two years would contribute to the development of curriculum units based on state standards in the core subject areas, would then use these units in their classes, and would regularly evaluate student achievement using state test scores and district assessments designed for grades not tested by the state. To carry out this professional development plan, the district decided to use the “trainer-of-trainers” model for all of its professional development efforts. One teacher from each grade at each school was designated a “lead teacher” who received special training. District staff consulted principals to select the lead teachers. Many of the lead teachers were group leaders during last year's process of aligning local curriculum and assessments to the state standards. These lead teachers would then supervise classroom teachers in grade-level groups from different schools in the development of the curriculum units.

During the summer of 1999, the district arranged for intensive professional development to be provided to all lead teachers by district and state specialists and university faculty. The professional development differed at each level, focusing on the district's priority areas. The lead teachers also received approximately 20 hours of training on adult learning, group process skills, effective staff development strategies, the use of data in decisionmaking, and team building to promote their effectiveness as teacher trainers.

Starting in September, the lead teachers conducted school-level sessions in grade-level groups (e.g., grades K–3, 4–6, 7–8, etc.) to assist all teachers in developing the curriculum units and using new teaching and assessment strategies to implement the standards. In addition, principals received a limited amount of staff development training to make them

---

<sup>2</sup> A fictitious school district.

aware of the new instructional approaches and to teach them how to encourage teachers to continue to learn about the standards.

At a district staff meeting, the superintendent announced that the Board of Education had requested evidence that the resources directed to professional development were leading to better teaching and learning. The superintendent felt this would be an opportunity for the district to explore the effects of its professional development. Based on his research, the superintendent said that districts that evaluate regularly for continuous improvement were more likely to effect long-lasting improvements in teaching and learning than those that do not. He asked Al Mitchell, the professional development coordinator, and Jane Evans, the curriculum specialist, to work together on a plan for a one-year evaluation. Although only a small amount of money would be available for the evaluation, he suggested that significantly more funding might become available in the future, including funding for a broader evaluation to extend over the next two years.

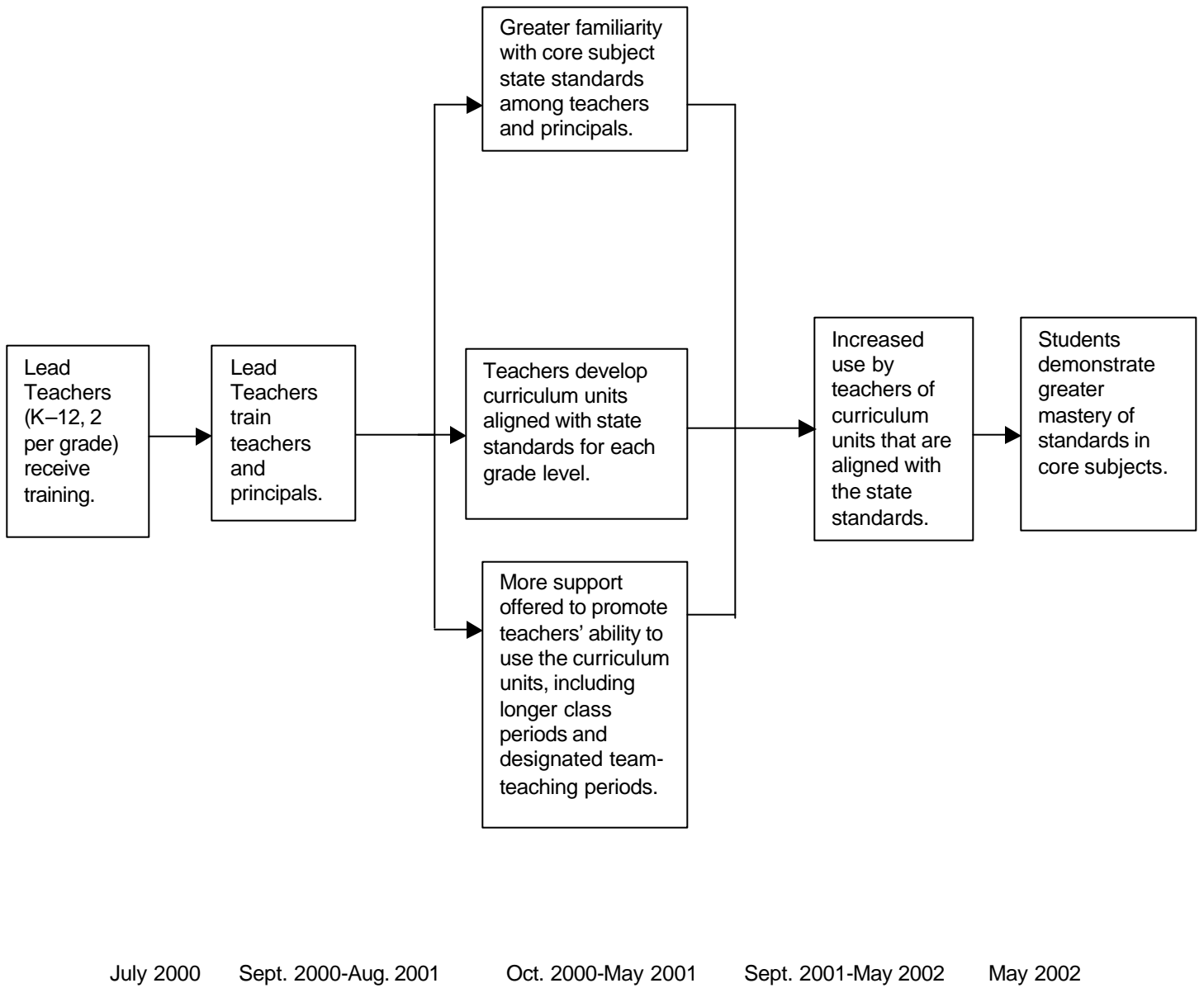
To decide on evaluation goals and objectives, Jane and Al created a logic model to describe the district program. They then shared this logic model with several of the lead teachers, who pointed out that the training of principals was not included on the model.

“It may not directly affect instruction, but then again, we don’t know,” said one. “What if a principal decides to arrange for additional time for teachers to meet? This kind of support ought to be included in your study, shouldn’t it?”

Al and Jane agreed and revised the logic model accordingly. Even if they didn’t specifically measure the effects of training on principals, they felt that anything a school principal did to affect the implementation process ought to be discussed in their evaluation. Exhibit 2 shows the logic model that they developed.

After reviewing their completed logic model, Al and Jane realized that it was unreasonable to look for effects on student achievement in this first-year evaluation. The professional development program hadn’t been in place for a sufficient time—according to program planners, lead teachers, and classroom teachers—to expect such changes. But even if they did not use student achievement data extensively in their first-year evaluation, Al and Jane wanted to “set the stage” for future evaluation linking professional development to student outcomes. They decided they would speak to their superintendent about this, hoping to strengthen their case for future funding of subsequent evaluation activities.

**Exhibit 3: The Ringwood School District's Logic Model**



## Chapter 4. Stakeholders: Deciding Who Will Come on the Trip

**F**ew evaluations, like a class field trip, can be conducted single-handedly. In most cases, it will be helpful for you to create an evaluation team, with several members, each of who can bring different skills to the evaluation.

### Creating an Evaluation Team

***It is always important to involve a cross-section of the various stakeholder groups, including teachers, principals, and district staff members.***

The place to begin your evaluation is with the creation of an evaluation team. You may need to inquire as to who has experience in evaluation, but ultimately, your team will be formed from the stakeholders who *care* most about the program being evaluated and who can contribute to the process. Examples of important stakeholders within the context of schools and professional development include the following:

- ***Those involved in program operations***, such as district and school administrators;
- ***Those served by or affected by the operations***, such as teachers;
- ***Primary users of the evaluation results***, such as school board members, administrators, parents, and other members of the community.

Sometimes it is also advisable to create a larger advisory group that includes community members. This advisory group would make recommendations to the “working” evaluation team. In addition, be sure to find out who will be responsible for making changes in the program based on the findings from your evaluation. This individual should be kept apprised of your progress and be a part of your first group of readers.

To conduct the evaluation, we suggest that you establish a team leader who will be most responsible for implementation. This person will likely have the most experience in evaluation or research, will be doing most of the substantive work, and will serve as the main contact for information about the evaluation.

You will also need to ensure that the team has the resources to work effectively. The team members must be able to find common time to meet and discuss the evaluation plans and progress (this may require the provision of “release time” to some staff members). The

team leader should organize the evaluation team meetings. At a minimum, you will want to convene an initial meeting of your full evaluation team to identify goals and objectives for the evaluation. The team should meet again to review a draft of the evaluation plan (which may have been developed by only a few team members). After the plan is approved, the larger group could meet at some regular interval to review evaluation progress and preliminary results as they emerge.

## Should You Use an Outside Evaluator?

Whether or not your district should use an external evaluator as a key member of your team depends on several factors, including the availability of resources and the capacity of district staff to conduct an evaluation. The main advantage of using an external evaluator is that the experience and expertise of a professional may improve the validity, reliability, and general quality of your evaluation. A third-party evaluation consultant can also bring greater objectivity to the evaluation and, in many cases, encourage respondents to speak more freely about the program being evaluated.

A disadvantage of using an external evaluator is that the district must find a person who is suitable—which can be difficult in isolated areas—and must then rely on that individual to get the evaluation done. Some districts may not wish to develop a relationship with an outsider, or they may have had negative experiences in the past and not wish to depend on anyone to get their evaluations completed.

For districts that have the necessary internal capacity, conducting an “in-house” evaluation can also have certain advantages. For one, carefully planning and implementing an evaluation using resources such as this handbook can increase the capacity of the district to conduct future evaluations, and to better use data for ongoing decisionmaking. Programs may also be better designed when a district has a clear understanding of how program activities are to be linked to program outcomes in an evaluation. Finally, costs may be greatly reduced through the use of in-house staff.

At the same time, evaluations conducted by district staff are sometimes put on the back burner for too long. Staff already burdened with many tasks find that they really do not have the time required to conduct a high-quality evaluation. A responsible external evaluator is less likely to fall short.

## Using an Outside Evaluator

How do you go about finding an outside evaluator? One option is to talk to staff at a local university or college to locate someone with the requisite skills and experience. Another option is to issue a notice of solicitation for professional services. You may also want to speak with staff at other nearby districts about whom they may have used as an external evaluator. Because no credentialing system exists for evaluators, you will want to gather information and references from your potential evaluators. Ask the prospective consultant to provide you with a written description of his or her qualifications for your project, and carefully review his or her previous experience, formal preparation, and evaluation philosophy to ensure a good match with your needs.

If you cannot find or afford a local evaluator to conduct the evaluation, you may want to consider hiring an outside consultant to help you get started and/or to review your plans for an in-house evaluation. Most evaluators will be willing to work with you on the design and methodology without actually conducting the evaluation. Another specific task for which you may want to use an outside evaluator is the development or review of data collection instruments that have been designed in-house. With their help, you can then focus your attention on implementing the evaluation and writing up your results.

If you decide to seek outside help, keep in mind that the relationship between an outside evaluator and program staff resembles that of an architect and the eventual homeowners: You need each other's assistance to do a good job. The outside evaluator's experience, like an architect's, means he or she will be able to help you design and implement a good evaluation. But the evaluator needs to know what you want. Consequently, you must develop the evaluation plan together.

## Involve Key Stakeholders in the Evaluation

To increase the likelihood that your results will be used to improve your professional development program, you may want to involve your most "powerful" stakeholders throughout various stages of the evaluation. You can

- Focus on areas of interest to these stakeholders, including those over which they have some control;
- Discuss how the stakeholders would make decisions based on evaluation results;
- Summarize results in format(s) that are useful to these stakeholders; and

- Continue to interact with the stakeholders after the evaluation ends, especially to promote the constructive use of the evaluation results.

## Know Your Constraints

Limited personnel, funding, and time frequently prevent districts from investing in program evaluation—despite the fact that evaluation is acknowledged as a helpful tool for improving programs.

Therefore, at the start of your evaluation, you should determine what resources you have available, including staff, time, and money, as well as any constraints imposed by the program operation (e.g., when key activities are scheduled). It is far better to be clear about these costs before you get started than to have to ask for additional funds to complete an evaluation once it is underway. Unlike professional development programs, an evaluation that is almost, but not fully, implemented is typically of little value.

## Creating an Evaluation Team in the Ringwood School District

This was the first year that Ringwood School District had provided money for evaluation of district professional development activities. Although the budget was small, both Al Mitchell, the professional development coordinator, and Jane Evans, the curriculum specialist, recognized that they had an opportunity to demonstrate the value of regular professional development evaluations to the district.

Given their budget and time constraints, Al and Jane knew their evaluation would have to be simple. But since they had little experience conducting evaluations, they felt they needed help. Al contacted the education school at a local university and eventually spoke to a faculty member who had conducted several program evaluations at nearby districts. This faculty member had even developed a survey to evaluate a staff development program for her research. She felt the survey could be modified slightly and used to evaluate some of Ringwood's programs.

The faculty member suggested that one of her graduate students, who had taken several research courses, assist Al and Jane in designing a plan for the evaluation of Ringwood's professional development program. This student would receive credit for doing this project as part of his coursework, and the faculty member would review their work. The evaluation they discussed would be simple enough that Al and Jane could conduct most of it themselves, with a small amount of administrative support.

Al and Jane decided they would create a team to develop the evaluation plan and review progress throughout the study. This team would consist of the two of them, the graduate student, two teachers from the grade levels that would become the focus of the evaluation, and the district's testing specialist. Al and Jane also intended to submit the evaluation plan to the superintendent for his review, and to brief the Board of Education in December on their preliminary results.

## Chapter 5. Evaluation Goals: Determining Your Destination

Once you understand your program fully, you are ready to plunge into creating your evaluation plan. As when planning a trip, you first have to know where you want to end up. Similarly, the next step in the evaluation planning process is to determine where you want to go with your evaluation. As a general rule, all evaluations should address the issues of greatest importance to the intended audience, and provide this information as efficiently as possible.

***A good evaluation plan anticipates the intended uses of the data, and at the same time is feasible and practical to implement.***

### Establishing Evaluation Goals and Objectives

What do you want your evaluation to accomplish? One way to think about this is to look at the types of questions you expect to answer when the evaluation is completed. As noted in chapter 2, these can include the following:

- ***Implementation questions*** concerning how the program was implemented. For example, how many teachers attended the training? Did coaches develop relationships with individual teachers? How did these coaching relationships vary (e.g., in terms of time and quality)? Did teachers develop new thematic units? Did principals complete classroom observations of the teachers selected for this program?
- ***Impact questions***, which explore whether the program had the intended effect on various target populations. For example, did teachers understand the new reading curriculum? Did they indicate that they were more likely to use more inquiry-based lessons in their science curriculum? Did students use technology more often in the classroom to learn geography? Did the instruction promoted by the professional development lead to changes in student achievement on state and/or district exams?

To decide what is important, you will want to turn to the key stakeholders you have included in your planning process. These individuals have a vested interest in the results of your evaluation. Obtaining their input, however, may yield wide-ranging demands. “Tell us whether our teachers are getting anything out of all those in-services,” someone may say. “And those teacher team meetings—are they leading to any changes?” Those who supervise the implementation of the program may insist on knowing what is working now,

and to use those results to make improvements, rather than waiting a year or more after the intervention to measure longer-term results. On the other hand, district or school administrators and the school board may focus more on the overall impact of the professional development program, asking, “Has the program contributed to an increase in student test scores?” and, “Is there any way to document the program’s effect on instruction?”

***You will need to prioritize the different demands placed on your evaluation.***

One way to prioritize your stakeholders’ information needs is to create and study a list of all of these needs. Ask yourself several questions as you review this list: Which questions do several stakeholder groups have in common? Which questions need to be answered to satisfy stakeholders who must make decisions about current and future professional development for teachers? Which questions can be answered now, and which require a longer-term evaluation? Ultimately, you will want to sequence these questions in order of importance to the district, considering the needs of your stakeholders and the current status of implementation of your program.

After prioritizing these questions, you will need to decide which questions your evaluation will seek to answer and link these questions to measurable outcomes. To develop these questions and the outcomes associated with them, we must begin thinking conceptually about the relationship between professional development and student learning outcomes, which are the ultimate target of standards-based reform.

## Examples of Questions for an Evaluation of Teacher Professional Development

Guskey (2000) suggests a useful way to think about different types of evaluation questions for teacher professional development programs. He identifies five “levels” of evaluation questions for professional development programs:

- ***What are the participants’ reactions?*** This is probably one of the most common types of questions asked, and the easiest to measure and analyze. Examples include several questions: “Was the time well spent?” “Did the material make sense?” “Were the activities meaningful and well-linked to standards-based reform?” “Was the instructor knowledgeable and helpful?” “Did participants learn the intended skills?” These types of data are often collected through end-of-session questionnaires.

- ***Did the participants gain knowledge or skills and/or change attitudes or beliefs?*** These short-term or intermediate program outcomes are related to the direct results of the staff development on the participants (e.g., teachers). These outcomes fall into three categories: cognitive, psychomotor, and affective—or “thinking,” “acting,” and “feeling.” Examples could include improved knowledge of content or performance standards, increased ability to implement a new curriculum or pedagogical approaches tied to standards, and higher expectations for *all* students. Key questions include: “What should teachers learn?” “How would teachers’ attitudes or beliefs change?” These outcomes can also be collected through questionnaires, to be administered at some time after the professional development has ended.
- ***Does the organization support the desired changes?*** Organizational factors—including school district or state policies and practices—can be critical to the success of a professional development activity, especially when these policies and/or practices are not aligned with the expected changes in teachers’ skills. For example, if standards-based professional development is meant to foster more cooperative learning in classrooms, practices that encourage student competition may complicate implementation of teaching strategies. Clearly, the policies and practices surrounding instruction, as well as the culture and politics of the school, affect student learning. As you reflect on the goals of your program and speak to participants and program developers, you may begin to develop a sense of the organizational factors that are critical to the success of your professional development program. Meeting minutes, formal policy changes, and changes in the allocation of resources are evidence of organizational support. Relevant questions include, “Did the organization support and facilitate the desired changes?” and, “Were sufficient resources made available?”
- ***Do participants change their behavior?*** Here we are interested in whether participants use their new knowledge, skills, or attitudes on the job; that is, did what participants learned make a difference in their classrooms? Such information can be obtained through interviews or survey questionnaires, or more directly through observation of teachers in their classrooms.
- ***Do students seem to be learning the material better?*** The bottom-line question for most educators is, “What was the effect on the students?” Examples could include comprehension of the scientific method, improved attendance, higher grades, and better group work. You will be asking, “What should students learn?” “What would they do differently if they had learned it?” and, “How have students’ attitudes or beliefs changed?” Measures of student learning can include indicators of student performance (e.g., tests, portfolio assessments), as well as affective changes measured through questionnaires and/or interviews (e.g., level of classroom engagement, positive learning attitudes, attendance, and behavior).

Guskey (2000) refers to these categories as “levels” to convey the idea that they can build on one another. For example, if participants in a professional development program do not gain any new knowledge, skills, and/or attitudes, then it is difficult to argue that subsequent changes in classroom practice occurred. Similarly, if teachers do gain new knowledge, skills, or attitudes but do not use them in the classroom, it is impossible to attribute changes in student learning to the program. At the same time, success at one level does not automatically lead to success at the next level. Thus, a comprehensive evaluation would measure effects at each level.

Ideally, in a full-scale evaluation, you would answer questions at most if not all of these levels. But resource constraints will limit the number of questions that can be answered and will often require some careful thought about how to plan out evaluations over a long time period.

## Deciding on the Destination: The Ringwood School District Experience

Once they had created their logic model, Al and Jane sat down with the graduate student in May to develop specific plans for their evaluation. They realized they had to get started quickly. In two months, the teacher leaders would receive their intensive training, and in the fall they would begin conducting professional development sessions at the schools.

To focus their efforts, Al and Jane considered what the district's key stakeholders would want to know about the professional development program. The superintendent's main concern was whether the professional development was contributing to improved student achievement. The president of the Board of Education told Jane that the board's interest in the district's professional development programs centered on costs. "What benefits can we expect, over how long a time period, and how much will the district have to pay for them?" he asked.

As professional development coordinator, Al needed to know whether the new approach to professional development would lead to greater or more rapid change in instruction than the traditional direct delivery of professional development that Ringwood had used in the past. He also hoped to learn whether teachers were comfortable developing curricular units under the supervision of specially trained lead teachers.

Jane said that the Office of Curriculum and Instruction wanted to know how quickly teachers could move toward more aligned instruction, and whether some teachers would learn more quickly or be more successful than others. "I'd also like to know more about what teachers would like to learn from this evaluation," Jane added. She reviewed responses to a teacher survey she had administered last year that probed teachers' attitudes toward the new curriculum and the district-wide priorities. "One general concern of teachers," she reported after her review, "was knowing how to blend what they already do well with the new curriculum and training. Teachers are concerned about how to make these choices."

The following table lists the information needs discussed by the stakeholders Al and Jane identified.

Key Stakeholders	What They Want To Know about Professional Development
Superintendent	Its effect on student achievement
Board of Education	Expected effects related to costs
Professional Development Office	(1) Effectiveness of the new approach in changing teacher practice and as compared with traditional approaches (2) Teachers' attitudes toward the new approach to professional development (3) How the approach could be improved
Office of Curriculum and Instruction	(1) Teachers' willingness to participate in creating curricular units (2) Whether high-quality curriculum units are completed (3) How quickly teachers implement new curricular units (4) Variation in teachers' success at implementation
Teachers	Strategies for implementation that will be most effective

Thinking about the needs of these stakeholders, Al and Jane drew up the following purposes for their evaluation:

- 1) To determine whether the new approach leads to changes in teacher practices and student achievement aligned with the standards.
- 2) To collect information on implementation of the new approach, including what is and is not working, whether curricular units are developed successfully, which curricular units teachers are being implemented most rapidly and successfully, and how teachers are using the new curricular units in the classroom.
- 3) To develop specific plans for future evaluation activities—specifically, how the district could link test scores to the professional development program.

Even so, Al and Jane knew they had to focus their efforts more narrowly, as they would not be able to meet all of these goals this year. The graduate student pointed out that the use of the lead teachers in Ringwood's professional development program was a significant change that could be evaluated separately.

Al and Jane decided that their best option was to look at how the lead-teacher system worked in two elementary schools. Because elementary school classrooms keep students

and teachers together, this focus would be conducive to eventually linking teachers and students and incorporating student test scores into future evaluation activities.

After talking further to the key stakeholders, Al and Jane decided they would try to answer the following questions:

***To better understand implementation:***

1. Did the lead teachers think that their training prepared them to work with classroom teachers on developing curricular units?
2. Did the lead teachers feel their training was well linked to the state standards?
3. How important were the principals' training sessions to the program?
4. How did classroom teachers respond to the trainer-of-trainers approach? What advantages and disadvantages did they think it had over traditional methods?
5. Were some teachers more satisfied with the training than others? Why?
6. Which teachers appeared to be responding most favorably to the professional development? Did their progress have anything to do with characteristics of individual lead teachers?
7. Were curricular units completed for each grade level? Do they appear to be of high quality?

***To look at teacher outcomes:***

8. How did classroom practices change at these schools, in both lead and regular teachers' classrooms?

***To plan future evaluation activities:***

9. How can test results be linked to professional development (i.e., what tests should be used and at what grades, with what degree of confidence can/should the link be made)?

Al and Jane examined their research questions and their logic model and realized that their primary unit of analysis for a more comprehensive evaluation would be the grade-level groups in which the school-level teacher training occurred. After all, their professional development program was designed to strengthen the leadership and cohesiveness of these grade-level groups, with groups of lead teachers (one for each grade) working together with their fellow teachers.

However, they lacked sufficient funding to conduct a very large evaluation, and most of their questions for this year's evaluation focused on implementation. In fact, all but one of their evaluation questions ("How did classroom practices change as a result of the

training?") could be answered through a formative evaluation. Because it was too soon to expect to see results of the professional development on student test scores, they felt that a formative evaluation, focused on a small number of grade-level groups and the teachers within them, would provide the information they needed most immediately and could inform their subsequent evaluation activities.

Al and Jane recognized that the superintendent and Board of Education wanted information that was not on their list. But they agreed to work closely on a plan that outlined ways to collect such information in the future and link it to the questions they could answer now. They felt that the superintendent would be supportive about this—although their first order of business would be to review their plan with him.

## Chapter 6. Evaluation Design: Plotting Your Course

**A**t this point, you have documented your program and established the goals of your evaluation (i.e., selected your “destination”). This included specifying the types of research questions you hope to answer. The next step is to “plot your course”—like planning the logistical details of your field trip—to ensure that you will reach your destination.

For evaluation, this next step involves the selection of a research design; that is, deciding how information will be gathered and used to answer your evaluation questions. As noted in chapter 4, evaluation methods should be selected within the constraints of available time and resources. No one set of methods is best for all circumstances, nor is one feasible within the context of all programs. The job of the evaluator is to select among the array of approaches those that are best suited for the specific evaluation. This is part science—knowing which design or method yields the most valid answers—and part art—deciding what is really practical to do within the particular program setting. The choices you make, however, have important implications for the evaluation cost, the validity of the resulting data, and the appropriate interpretations of the findings.

### Choosing a Design Strategy—Things Evaluators Worry About

Can I Say that the Program Causes the Change?

Evaluators pay a good deal of attention to determining whether a program caused some outcome or result. For example, you may want to know whether your professional development program *caused* teachers to change their instructional practices, and if so, whether these classroom changes, in turn, *caused* an increase in student achievement. This determination of **causality**, according to Cook and Campbell (1979), requires that three conditions be met:

- **Temporal Order.** Not surprisingly, the expected cause should occur *before* the expected effect. If a tree fell before the car was in the vicinity, we would not conclude that the car hit the tree and caused it to fall.
- **Covariation.** Changes in the expected “cause” must be related to changes in the expected “effect.” For example, if you vary the amount of training provided to teachers, you should observe changes in the outcome measure, whether it is measured in terms of the teachers’ degree of implementation of the newly acquired skills, or the test scores of students in the classrooms of teachers who were trained. In other words, *more* training should be associated with *greater* knowledge and/or skills.
- **No Plausible Alternative Explanation.** If other possible causes for the observed effects exist, then you cannot be confident that your professional development program caused the changes that you observe or measure. For example, if student test preparation was a major emphasis in the district, you cannot say with confidence that the professional development program *caused* the increase in student achievement. It may have been better test-taking skills.

For the most part, these criteria match how most of us think about causation in our everyday lives. But from the perspective of designing an evaluation, it is the third condition that is the most difficult to satisfy; that is, being able to eliminate other explanations for why certain outcomes or changes were observed. As discussed in chapter 1, districts that engage in systemic reform typically make several changes in their education system, often simultaneously. As a consequence, most of the work of crafting a good evaluation design will involve eliminating or reducing possible “threats” to the validity of your conclusions about a program’s effect. The collection of contextual information about the program, as discussed in chapter 3, helps to identify other plausible explanations.

#### Are the Conclusions Valid?

Another important consideration in any evaluation is the validity of the conclusions that can be drawn from the data. This means that your conclusions should be justifiable, relevant, meaningful, and logically correct. Evaluators use two technical terms when discussing validity:

- **Internal validity** is the extent to which one can claim that the program made a difference; that is, “If test scores go up after we started the teacher training program, can the changes be attributed to what we did?” In other words, can you claim that the professional development program *caused* the observed improvements in test scores?
- **External validity**, in contrast, is about the ability to generalize beyond the single study to other participants and similar programs; that is, “If professional development changed the classroom practices of the initially trained teachers, can we expect the same results for all of the teachers we train? In other words, are the results *representative* of all teachers and school settings?”

A strong evaluation design must have high internal validity, but not necessarily high external validity. That is, you always want to ensure that you have a strong basis for drawing conclusions about the program or target group of participants (e.g., teachers) you studied. Whether external validity is important will have to be determined within the goals and objectives of your evaluation.

#### What Can Affect Validity?

What can affect your ability to attribute effects to your professional development program?

Below are some possibilities:

- **History.** Here the argument is that it is not the professional development that caused test scores to increase, but something else that happened during the same time, such as the simultaneous introduction of a new curriculum or a special test preparation program implemented by the district.
- **Maturation.** One can also argue that changes in student test scores (e.g., from 4<sup>th</sup> to 5<sup>th</sup> grade, or from fall to spring) would have gone up regardless of the program because of children's normal maturation or growth .
- **Testing.** In some cases, how students do on a test the first time they are tested (what is referred to as the "pre-test") can affect the students' scores on a subsequent test (called the "post-test") regardless of the intervening program. This is a statistical concept that is easiest to explain by example. If, for example, one were to test a group of low-performing students at two points in time, one would observe an increase in test scores even if you did *nothing* to them. This is because by chance alone some of the lowest scorers will "regress to the mean," or have higher scores at the second measurement point.
- **Instrumentation.** In some cases, evaluators may be forced to use a different tool to measure conditions before and after the program, and this change can explain any observed differences. For example, the district may have changed the test used to assess student achievement, so the data available on students before the teacher training is based on a different test than the data that are available after the training has been completed.
- **Dropouts.** In some training programs, the less-skilled teachers may "drop out," leaving only the more motivated and initially skilled teachers around for the post-test. As a consequence, average scores on a test of teacher competencies administered after the completion of training could be higher than those recorded before the training, not because of the effect of the program but because of changes in who was tested.
- **Selection Bias.** Similarly, one may be comparing outcomes for teachers who received training to those who did not, and the two groups may not have been comparable at the beginning. For example, the more motivated teachers may "volunteer" for the professional development. The extent to which the two groups being compared (teachers who did and did not receive training) are not similar at the beginning of the evaluation is related to what is called "selection bias." Selection effects can also occur during the evaluation if any of the threats described above affect one group differently than the other. For example, if one group was exposed to

different changes in their school, or was treated differently in terms of data collection, or if there was a differential rate of drop-out from the study between the two groups, selection bias exists.

***For the most part, the challenge in evaluation research design is to minimize the effect of various threats to validity.***

The remainder of this chapter will examine the different types of research designs. Subsequent chapters will focus on data collection, data analysis, and reporting of evaluation results.

## Using Real-World Examples to Understand Evaluation Design

To help ground this discussion in real-life examples, we have taken the liberty of creating two dimensions along which teacher professional development programs can vary. This is an obvious simplification of a complex field, and we use it only as a device to aid the presentation of complex technical details:

- ***The Focus of Professional Development.*** First, one can think about professional development strategies as being defined along a dimension that we will call the ***focus***, or ***“unit of analysis.”*** On the one hand, strategies can be ***teacher-focused***—this would include sending teachers to external classes or providing district-sponsored workshops, seminars, or summer institutes. The individual teacher would be the focus of the training and the “unit of analysis” for the evaluation. On the other hand, a district may choose to adopt ***school-focused*** strategies that involve the entire staff at one or more schools or entire grade- or subject-level teams. This type of strategy could include coaching/mentoring programs, developing in-school “learning communities,” and creating teams of teachers to develop curricula aligned to standards. The point here is that a collection of staff, rather than individual teachers, is the target of the intervention.
- ***The Scale of Professional Development: Saturation vs. Sample.*** In addition to varying the focus of the professional development, a district can decide to either train all the eligible staff—a “saturation” model—or to train subgroups of the eligible staff, such as by selecting certain schools or staggering the implementation of the training over time.

In addition to these two program dimensions, your evaluation can focus primarily on formative or process evaluation questions, or it can focus on summative or impact evaluation questions. Because the design of a process evaluation is, for the most part, unaffected by the two dimensions listed above, we will first examine research designs for this type of study, and then discuss impact evaluations within the context of the two dimensions of professional development described above.

## Formative or Process Evaluation Design

***Formative evaluations explore whether the program has been implemented as planned—and how well. Questions include who is being served, to what extent, how, and at what cost. When you conduct a formative evaluation, you will want to include all of the treatments, or program design elements, in your research.***

Formative evaluations of newly implemented programs are often conducted to help staff understand their progress and make midcourse improvements. For example, in many districts, professional development programs are implemented in ways that prevent teachers from making the changes needed to eventually affect student achievement. Formative studies can help determine how well the process of teacher training is proceeding, and can identify obstacles to implementation before they threaten the effectiveness of the program.

Formative evaluation generally imposes fewer demands on the design of a study than summative evaluation because there is no intent to attribute changes in outcomes to program activities. For example, one of the common objectives of formative evaluation is to compare expectations with actual performance; that is, is the program being implemented as planned? Did the proposed activities occur as expected? Were the intended staff involved? Were any school-level changes necessary to support the implementation of professional development, such as the use of incentives to encourage teachers to participate? Were these provided? Did they work as expected? Was any follow-up activity planned as part of the implementation process? If so, did it occur? What resources were planned to implement each component of the program, including planning, training, materials, and evaluation? What were the actual costs?

Typically, these questions can be answered by using program files (e.g., attendance sheets, office records), simple surveys, and sometimes interviews and focus groups. You may also wish to observe some of the professional development sessions to confirm that the program is being implemented as planned. For example, if trainers are not covering the expected material or giving teachers the opportunities to learn it, then there is no way that teachers will change their practice in the anticipated manner.

Formative evaluations often depend more on qualitative than quantitative data. Each type of data has its own unique strengths, however:

- **Qualitative data**, such as data gained from in-depth interviews with a small sample of voters, provide rich information, depth, and a focus on specific (often limited) populations. Qualitative data depend on the skill of the researcher, although recent techniques have helped make these types of methods more rigorous.
- **Quantitative data** are generally better for capturing the breadth of a program and its participants in a way that is generalizable, such as through the use of a national polling survey that goes out to thousands of voting-age adults. Quantitative data are typically collected through standardized methods that are often believed to yield more scientifically rigorous results.

Some researchers also argue about the philosophical underpinnings of the two research methods. For example, qualitative researchers argue that there is no “absolute truth” and that all knowledge is “constructed,” so there is no a priori advantage to quantitative methods. Quantitative researchers, while unlikely to claim absolute truth, point to the greater representativeness of this method.

Another job of formative evaluation is to assess **how well** each step—from the initial professional development session to implementation of the training in the classroom—was implemented. For example, in the “trainer-of-trainers” approach, did the lead teachers gain the necessary knowledge and skills from their training to serve as trainers? Did they understand their roles as trainers? Did they perceive, after meeting with the classroom teachers, that they had provided the mentees with new information that the mentees could use? You may need to interview classroom teachers to see if their views corroborate the lead teachers’. Did classroom teachers feel that they were prepared to implement new practices in their classrooms? Assessing whether and how well each of these steps has been implemented, including the proposed timeline, will be the primary activity for your formative evaluation.

In addition, you will want to check on several factors that may be promoting or inhibiting the implementation of your professional development program. For example, you may want to determine the following:

- The participants’ satisfaction with the professional development experience, including location, environment, and other seemingly small factors that could have a significant effect on their engagement in the training;
- What factors participants perceive as facilitating implementation (e.g., principals’ support, materials provided at session);
- What factors participants perceive as inhibiting implementation (e.g., lack of time to explore new resources, student population challenges);
- How participants feel the program could be improved; and

- The extent of progress that providers and participants perceive has been made toward achieving the goals of the program.

This last point—checking on progress toward program goals—can give you a sense of how effective the program appears to be so far. You will of course need to identify knowledgeable and representative data sources to assess such progress. In many cases, your data sources will be people—some or all of the teachers targeted for participation, the program providers, the principal, or others involved with the program. You may choose to collect these data systematically, such through a short survey of all participants, or more informally, through interviews with program providers and a small number of participants.

In addition to collecting self-reported data, you may want to observe one or more professional development sessions, if feasible, to validate reported information. For example, a lead teacher may report that classroom teachers were more engaged in the professional development session than they actually were. You may also observe problems in the implementation that can be noted and used to improve the program. Observations of teachers' implementation of staff development training in the classroom are also very useful. In the case of a trainer-of-trainers model, you might even be able to use your lead teachers to conduct observations of classroom teachers with whom they have not worked, thereby minimizing bias and enhancing lead teachers' knowledge of classroom practices in general.

***In formative evaluations, it is usually important to observe the context before the program begins, and at least once after it has been implemented.***

Critical to any evaluation is the collection of baseline information, which allows you to understand what, in fact, has changed in the program (comparing before and after). If you can examine the program at multiple points during its implementation, you can also investigate the extent to which changes have been made. You will also want to compare the observations of the program to what was intended; that is, comparing program plans to actual program implementation.

***Process evaluations also play an important role in summative or impact evaluations.***

In addition to being “stand-alone” evaluations, process studies are often included as part of impact evaluations to help evaluators understand the nature of the actual “treatment”

being studied and to provide a deeper understanding of why certain program effects are, or are not, observed. Documenting and reflecting on how the different program elements interact will allow you to more legitimately link the delivery of professional development to your expected outcomes; that is, you will be better able to eliminate other plausible explanations.

## Applying Formative Evaluation to Teacher Professional Development: The Ringwood Experience Continued

Al and Jane decided that their formative evaluation would have two parts: first, a determination of whether the program was being implemented as planned, and second, an in-depth look at how the program was proceeding in two elementary schools.

To assess whether the program was being implemented as planned, Al and Jane reviewed the specific components of the program for the entire district:

- Training of 84 lead teachers (7 from each of 12 elementary schools)—during the summer.
- Training of principals—two workshops during the school year.
- Training of regular teachers—to occur in grade-level sessions conducted at the schools by lead teachers. The *minimum* treatment, according to the professional development plan, was four sessions every half-year.

Al and Jane reviewed program files to determine what they already knew about some of the training sessions—specifically, whether all of the lead teachers had attended the training, and how much money was spent on the lead teacher and principal training sessions. They lacked records about the training conducted by lead teachers at the school sites, but knew the lead teachers kept such records to account for their own and their colleagues' time. To gather this information systematically, Al and Jane decided to administer a short survey to lead teachers, with the following questions:

- How many training sessions of principals did you conduct at your site? Who attended? How long did each training last? How long did it take you to prepare and conduct these training sessions? What additional costs, if any, were incurred?
- How many training sessions of regular teachers did you conduct? Who was there? How long did each training last? What additional costs, if any, were incurred?

Al and Jane designed the survey themselves. It was short and would be easy to administer. Using the responses they would collect, as well as their program files, Al and Jane felt confident that they could determine how fully the program was being

implemented at this time, as well as the amount of resources allocated to each of these program components.

To understand the treatment better, Al and Jane reviewed the evaluation questions they had developed earlier. For each question, they identified sources for questioning (indicated here in parentheses):

- Did the lead teachers think that their training had prepared them to work with classroom teachers on developing curricular units? (lead teachers)
- Did the lead teachers think their training integrated this work with the state standards? (lead teachers, review of training materials)
- How important were the principal sessions to the program? (lead teachers, principals)
- How were classroom teachers responding to the trainer-of-trainers model and to their participation in curriculum development? What advantages and disadvantages did they think it had over traditional methods? (lead teachers, classroom teachers, principals)
- Were some teachers more satisfied than others? Why? (regular teachers)
- Which teachers appeared to be responding most favorably to the professional development? Did their progress have anything to do with characteristics of individual lead teachers? (lead teachers, regular teachers, principals)
- How much progress was made in developing curriculum units, and what were teachers' impressions of the quality of the products so far?

Al and Jane recognized that evaluating changes in classroom practice as a result of the professional development might not be possible yet and that linking the professional development to student test results definitely could not be accomplished given the time frame for this first evaluation. However, Al and Jane did discuss talking with the lead and classroom teachers about these issues during the evaluation.

To answer these research questions, Al and Jane decided to focus on two schools, and to focus on both grade groups (K–3 and 4–6) at each school. They designed interview protocols to answer all of the questions, using simple scales to gather information for each question. They also observed one school-level training session, conducted by lead teachers, at each school, and reviewed the lead teachers' preparatory materials for the sessions, to get a sense of the different approaches being used.

Al and Jane conducted the interviews themselves, and then had an administrative assistant compile all of the responses into an Excel database. The quantitative data

(responses on scales) were analyzed using the database, and the open-ended responses were compiled and analyzed by Al and Jane, separately. They then came together, reviewed all of their results, and wrote a five-page report on what they had learned.

## Types of Impact Evaluation

Impact evaluations, unlike process studies, are intended to link outcomes to the activities of the program, and this requirement significantly complicates the design of sound evaluations. In general, impact evaluations can be grouped into three categories:

- **Randomized or experimental designs.** The first type of design is the true experimental study in which participants (teachers) are randomly assigned to receive either the treatment (the professional development program) or to receive no services (referred to as the control or comparison group) and generally continue with “life as usual.”
- **Quasi-experimental designs.** The second category of impact evaluations generally includes comparison groups and/or multiple measurement points, but individuals are not randomly assigned to study groups. That is, the evaluator does not exercise control over the creation of the treatment and comparison groups.
- **Non-experimental designs.** The final—and weakest—type of impact evaluation excludes the use of a comparison group. The simplest example of this approach is a single-shot survey used to describe the characteristics of a group of individuals or programs (e.g., a survey of the teaching practices of teachers in the district).

As shown in Exhibit 4, **randomized experiments** are typically considered to be the strongest approach to impact evaluation, and are often the “gold standard” against which other impact evaluation designs are judged. Each of these approaches to impact evaluation is discussed below.

**Exhibit 4: Types of Impact Evaluation Designs**

<b>Evaluation Strategy</b>	<b>Researcher control over program treatment?</b>	<b>Intervention intended to suit research design?</b>	<b>Causal interpretability?</b>	<b>Generalizability?</b>	<b>Cost</b>
Randomized Experiment	Yes	Yes	True causal inference	Limited to test sites, treatments, and population	High
Quasi-experimental	No	Yes	Possible causal inference	Limited to test sites, treatment, and populations	Moderate
Non-experimental	No	No	Weak inference	Can cover wider range of sites and populations	Moderate to Low

Adopted from Frechtling, J., editor. 1995. *Footprints: Strategies for Non-Traditional Program Evaluation*. Arlington, VA: National Science Foundation.

### Randomized Experiments of Professional Development

***In a randomized experiment, individuals are randomly assigned (e.g., by the “toss of a coin”) to either a treatment (or program) group that receives the services under investigation, or to a control (or comparison) group that does not receive the services.***

The process of using random chance to decide who gets the services ensures that the two groups (treatment and control) are equivalent at the start. Assuming that you collect comparable data from all members of both groups, a comparison of the average outcomes (e.g., teacher competencies) across the two groups will yield an “unbiased,” or very accurate, estimate of the effect of your professional development program.

For example, an experiment might involve randomly assigning half of your teachers to participate in a professional development program (e.g., a special summer institute focusing on teaching new academic standards in math), and assigning the other half to a control group (i.e., those who will *not* attend the training). If you then observed differences in the subsequent teaching practices of the two groups, this would be a highly valid measure of the effect of the training program. That is, you could claim with confidence that the training *caused* the observed change in pedagogy.

One common objection to experiments of this type is the denial of services to those assigned to the control or comparison group. One option for addressing such concerns is to stagger the treatment, with some individuals being randomly selected to receive the treatment now and a second group (who will serve as the comparison for the first group of participants) to receive the treatment later. This approach can be very useful when programs do not have the resources to serve everyone at the same time, or when everyone is not required to receive the same set of services. This model, however, allows for a determination of only short-term effects, since everyone will receive the treatment eventually.

Another point to be made about experiments is that although one can use only post-test scores to estimate the program's effect (if random assignment has been properly implemented), it is a good idea to collect baseline information (i.e., at a time prior to the provision of program services) to examine the comparability of the two groups. These data can also be used later in your analysis to increase the reliability of the estimated program effects.

## Applying Randomized Experiments to Teacher Professional Development: Two Examples

### Example 1: Elton School District

In Elton, Ellen Reams, the professional development coordinator, was responsible for evaluating a new English language arts program aligned with state and district standards, to be offered to high school teachers during the upcoming summer. Typically, the district exercised authority over a portion of teacher professional development, as when all teachers in the district were required to participate in technology training last year. Thus, Ellen had some control over who would participate in the new staff development program.

After consulting with her superintendent, Ellen was given permission to randomly assign half of the district's high school teachers to the new staff development program, providing her with a treatment and a control group and the ability to make legitimate comparisons about the two groups. However, she recognized that some of her treatment- and control-group teachers might participate in other professional development that could affect their teaching. Although a random assignment ensures some level of equivalency among treatment and control groups, she decided her evaluation would be improved if she "controlled" for this variation as fully as possible. This required that she gather information about what *additional* professional development teachers received.

To conduct her evaluation, Ellen administered a survey at the end of the school year, before the professional development was offered. The survey probed secondary language arts teachers' attitudes and current classroom instructional practices. A follow-up survey would be offered in June of the following school year, allowing 12 months for teachers to make changes in their instructional practices. This survey included questions about the type and extent of professional development in which her treatment and control teachers had participated.

#### Example 2: Mayberry School District

Mayberry School District, a small district, presents a greater challenge to the use of a randomized experiment. Joel Sanders, the curriculum specialist, is about to institute a mandatory collaborative study program for *all* teachers in the district. After presenting teachers with model performance assessments designed to assess student mastery of the standards, the teachers will be required to review these assessments and develop their own tests for the grade level and/or subject they teach. They will then reflect together on their progress in teaching with these performance assessments. The goal in Mayberry is to foster the creation and use of learning communities at all schools and among all teachers. These learning communities are expected to increase teachers' understanding of how to improve student performance through the use of assessment data. The district hopes that this activity will also lead to more collaboration among teachers on how to promote and assess student achievement.

The challenge for Joel was how to evaluate a program that would "saturate" the entire district. He did not have the option of assigning some schools to the program and leaving others out, since he and other district officials felt each school needed to be involved in this important process. However, Joel suggested to his colleagues that not all schools had to begin the process simultaneously. Some schools could begin in the fall, while others could begin the following year, and these decisions could be made on a random basis. Another option was to start the collaborative study groups in all schools, but to begin this year in certain grade levels at some schools, and in other grade levels at other schools. The advantage of this second approach is that all principals would be trained at the same time (as their support was a part of the program), and schools could move more gradually to the full model, which requires additional planning time and other organizational changes that can prove difficult to implement all at one time and schoolwide.

Joel and his colleagues decided on the second option. In the fall, they would institute the collaborative study program in the upper grades at half of the schools, and in the lower grades at the other half of the schools. Schools would be randomly assigned into upper or lower groups, allowing comparisons between the same grade levels at treatment and control schools. Baseline and follow-up surveys would be administered over the course of the upcoming school year. The district would study the results over the summer and, Joel hoped, would then implement the model fully, in all grades in all schools, the following year.

## Quasi-Experimental Impact Studies

***The second category of impact evaluations, quasi-experimental designs, lacks the random assignment of eligible program participants to either a treatment or control group, and instead uses other methods to construct a group that is “similar to” the treatment group and that can serve as a basis for later comparison of outcomes.***

A wide range of quasi-experimental designs exists, but most fall into two broad classes:

- **Assessment of Treatment Group Only.** Probably the most common type of quasi-experimental evaluation relies on the collection of data only from program participants before and after they receive the intended services. For example, one might assess teacher knowledge and skills (or the achievement of teachers' students) before and after the teachers receive training. In effect, the individual teacher serves as his/her own “control” group. That is, characteristics measured before the training is delivered are the basis for judging whether there were any changes after the training was completed.

Such designs are relatively inexpensive and easy to implement, which is why they are so common. But their ability to attribute differences to the intervention is generally weak. For example, suppose students' reading test scores went up after teachers participated in a training program. Could you then conclude, with certainty, that the professional development activity caused this gain in student achievement? Probably not. As noted above, students learn through a variety of mechanisms, and their normal rate of maturation would lead to some gains even without a better-trained teacher. As a consequence, without some information about what would have happened to them ***in the absence of the teacher-training program***, you really cannot be certain about whether the program had an impact on student achievement.

One way to establish linkages between programs and student results might be to compare the gains to some standard, such as averages for students in different grades, or from test norm data in the case of a standardized test. But these approaches may be misleading, especially if the treatment group is different from the individuals used to create the norm.

- **Assessment of Treatment and Comparison Groups.** A more rigorous way to determine the effect of professional development without using a randomized experiment is to compare the performance of the teachers who received training with their peers who did not participate in the training program. Such individuals form what

is called a “comparison group” (strictly defined as distinct from a control group, for which teachers are randomly assigned). Their characteristics—particularly the outcomes of interest—are measured at the same points in time as those of the treatment group. That is, if “before” and “after” information is collected from both groups, then a comparison of the differences (i.e., the pre-test/post-test difference) represents a measure of the program’s impact.

If you cannot implement a true experiment, you should include, if at all possible, a comparison group of teachers in your evaluation. This will greatly increase the confidence that you will be able to attach to any conclusions you draw from the completed evaluation. Options exist for improving your quasi-experimental study, and some of the more common methods are provided in the text box below.

- **By Expanding across Time.** In addition to adding a comparison group, you can also add multiple measurement points. These can be “pre-tests,” which occur before the treatment is received (also called “baseline” measures), and “post-tests,” which occur after the participants have completed the program. Several pre- or post-tests can be added with more measurements to further strengthen the design. Pre-tests allow you to control for any existing differences between those who do and do not receive the treatment, and having multiple measurement points allows you to control for differences in the rates of normal maturation (with multiple pre-tests), or to examine the extent to which gains “fade out” over time (with multiple post-tests).
- **By Expanding across Treatments.** Another way to improve your quasi-experimental design is to administer multiple treatments that vary in approach or intensity. For example, you could test differences in formal versus informal professional development, or in different amounts of training, by assigning teachers to different treatment programs.
- **By Expanding across Groups** A final option is to segment different types of participants—new versus more experienced teachers—and assign them to separate treatment groups. This allows for a direct test of the extent to which outcomes may differ between the groups.

### Applying Quasi-Experimental Designs to Teacher Professional Development: The Kramer School District Experience

The Kramer School District has decided to embark on a multi-year effort to upgrade the skills of their instructional staff to ensure that all teachers are equipped to help students attain the state’s new academic standards. In addition, the district’s students have not been doing well on the annual state assessment, and there is, not surprisingly, great pressure to show significant improvement.

The staff development office has been collecting information on a variety of models of professional development and has selected two that they believe show the most promise for their teachers: creating summer institutes that would bring teachers together for two weeks to work on the development of curriculum and lesson plans linked to the state standards and assessments; and creating mentoring relationships within schools to team master teachers with newer or less skilled teachers. Because these are both relatively expensive options—given the total number of schools and teachers in the Kramer District—the staff development office has received approval from the School Board to “pilot test” both options before going to scale across the entire district.

When the staff sat down to plan their evaluation of the initial pilot test, they realized that they had two distinct treatment options:

- **teacher-focused** design, using the summer institutes, and
- **school-focused** design, involving groups of teachers in the same school.

Because both were being implemented on a pilot basis, the staff knew they could find similar teachers and schools who could serve as comparison groups for their study (i.e., they were not planning to saturate the entire district). They knew they needed a comparison group because they wanted to determine whether the teachers who participated in both programs improved both their knowledge of the standards and what was needed to get students to those standards and in their actual classroom teaching behavior.

At one of their early planning meetings, one of the team members had a great insight: “Because we have two different initiatives, we can test the benefits of three rather than two options.” Here’s how:

- **Summer institutes alone**—teachers who attend the summer institute who are *not* in a school with the mentoring program;
- **Mentoring program alone**—teachers in a school with a mentoring program who do *not* attend the summer institute; and
- **The combination of the summer institute and the mentoring program**—teachers in a school with a mentoring program who also participate in the summer institute.

“All we need is a single ‘comparison’ to represent the group with *neither* form of professional development!” concluded the team members.

They decided to focus on 3rd- and 4th-grade elementary school teachers for their pilot test. This represented about 380 teachers in 60 elementary schools. They also decided to implement the mentoring program in 20 schools, and to restrict the first summer institute to about 80 teachers. They planned to have one master teacher in each of the selected mentoring schools who would each work with three teachers (a total of 20 mentors and 60 mentee teachers). They would select the mentoring schools by “matching” all of their elementary schools so that each one had a comparison school. This was done using available information on student characteristics and prior scores on the state assessments (which are given annually to students in all grade levels).

Their plan then looked like this:

- Twenty schools would have a mentoring program involving 80 teachers. Half of these teachers would be invited to participate in the summer institute (i.e., half would receive only the mentoring program and half would receive both programs).
- The remaining 40 teachers for the summer institute would come from the 20 matched comparison schools.
- Teachers in the 40 schools that were *not* involved in either training component would serve as the comparison group for the evaluation.

The staff would use self-administered surveys to collect information about each teacher’s knowledge and attitudes about teaching in general and about the state standards, both before the program began (the pre-test), at the end of the school year (first post-test), and at the start of the next school year (to capture the summer institute effect; the second post-test). For a subsample of teachers in all of the study groups, the staff development office would also conduct classroom observations. Finally, information on the test scores of their students would also be acquired from district records. The same data collection would be done for both the teachers who participated in one or more of the training components, and for the comparison group teachers as well.

We will revisit this example in later chapters to illustrate issues related to selecting samples, designing data collection, and analyzing the data.

## Non-Experimental Impact Studies

**Non-experimental evaluations should be avoided as a basis for estimating program impact, as they will not yield convincing results!**

The third type of impact evaluation designs, called *non-experimental* evaluations, encompasses a variety of methods. The simplest evaluation strategy is what is called the “single-group, post-test-only” design. As the name suggests, in this approach you would study a single group, comprised of individuals who receive the treatment, such as a group of teachers who are selected to receive professional development. These individuals would be observed once *after* they have completed the program (hence the name “single post-test”). This is a particularly weak research design because there is no measure of where the participants started (e.g., the skills that teachers had before the training program), and you do not know if your group of participating teachers is representative of typical participants (e.g., did only the most motivated teachers take the training?).

### Choosing an Impact Evaluation Design

How do you choose a design from all of these options? Debates that rage between strategists within the evaluation profession generally have each claiming the superiority of their position. In reality, most good evaluators are familiar with the different design strategies and use them in various combinations as the need arises. In recent years, increasing attention has turned to how one might integrate results from evaluations that use different strategies, carried out from different perspectives, and using different methods. Clearly, there are no simple answers here. The problems are complex, and the methodologies needed will and should be varied.

Regardless of the approach used—and it is often recommended that a combination of methods be used—evaluations must be guided by certain practical considerations:



### **Practical Guidelines for Evaluation**

1. The results must be credible to the stakeholders; most evaluators face a skeptical audience for their work, and the study must be viewed as credible if the results are to be accepted.
2. The organization must have the staff skills necessary to carry out the study.
3. The evaluators have to work within the constraints of available money and time. In many cases, these resources alone can determine the evaluation design that can be used.
4. The evaluation design should use multiple methods. This allows for the triangulation of research findings, strengthening the validity and credibility of the results with multiple pieces of evidence, derived from different sources that point in the same direction.

## Chapter 7. Data Collection: Getting Started

**S**o, you have now planned your destination for the field trip and decided how to get there. Now it's time to figure out what information you want to collect, where to get it, and how to get it. Consider this the final planning for your field trip!

### Deciding What Information to Collect

***The place to begin making decisions about what information to collect is with your evaluation goals and the research questions you have already developed.***

To ensure that you collect the information you need by the end of your evaluation, you will want to identify the data you need for each of your evaluation goals and research questions (in the next chapter we will discuss linking these to information sources).

To help you do this, you should construct a table that links research questions to measures and indicators that will next be used to develop actual questionnaires and other data collection instruments. An example of such a table is shown below in exhibit 4, using the example of the Kramer School District discussed at the end of chapter 6.

As shown in this table, you will want to begin in the first column by writing down the research questions that you used to develop the design for your evaluation. Next, you need to specify the types of data that you will need to answer the question. For example, if your research question is, "Are teachers who have received mentoring on how to teach the state's new academic standards more likely to use the new pedagogical approaches in their math classrooms?" then examples of this measure might include "the incidence of using new instructional strategies in math instruction."

***Exhibit 5: Linking Research Questions, Measures, and Data Sources***

**Exhibit is in separate file.**

Once you have specified all the data you will need to answer your research questions, you will need to determine how you will obtain each data element. As will be discussed in chapters 8 and 9, this includes both the source (e.g., teachers, principals, students), and how the data will be collected, such as through in-person interviews, classroom observations, or self-completed questionnaires (you may have multiple sources of information for the same measure or indicator, and this can be good if it allows you to either contrast/compare information from different respondents or use information from different sources as a way to identify the most accurate data, a process researchers call “triangulation”). Finally, the last two columns of this table would be used to document the specific data collection instrument (and question number) that links to each data element, and provides space for any comments that you may want to record for later reference (e.g., these data to be collected only at the pre-test).

### To Sample or Not To Sample, That Is the Question

***If you cannot collect information about everyone who may be involved with, or affected by, your professional development program, you will have to make decisions about whom to include in your evaluation.***

In many instances, it is infeasible or too costly to collect information from all of the teachers or schools that may be part of your professional development program. In such situations it is appropriate to select a sample or subset of the total number of program participants. Of course, if the program is only being implemented on a small scale, or if your district is small enough, it may be possible to include everyone. If you do not have to sample—that is, if you can collect data about *all* participants—that is almost always a better approach, as you will eliminate one source of error that plagues many studies, what statisticians call “sampling error.” The term describes the inaccuracy that is introduced because one has only observed a subset of all possible individuals or programs.

But let us assume that for one reason or another you cannot include everyone. How do you decide which individuals or schools to include in your evaluation? There are two general ways that you can make these decisions: through either ***probability sampling*** or using some type of ***non-random*** selection procedure.

## Probability Sampling

***Probability sampling involves the use of statistical procedures to select study members in a way that gives each eligible participant a known chance (or probability) of being selected.***

In probability sampling, the possibility (or chance) that a particular school or teacher is selected for your study does not have to be the same (i.e., equal), it merely has to be known to you and the members of your evaluation team. For example, if you had 10 schools and randomly selected two schools to include in your evaluation, every school would have the same 0.2 chance of being selected (i.e.,  $2 \div 10 = 0.2$ ; this can also be stated as odds, like in a lottery; that is, each school has 2:10 or 1 in 5 chance of being selected).

The place to begin the process of probability sampling is to ask yourself, "To whom do we want to generalize?" In most social research we are interested in more than just the people who directly participate in our study. We would like to be able to talk in more general terms; that is, to be able to say how this professional development program would affect all of our teachers or all of our schools.

***The group you wish to generalize to is referred to as the study population.***

The study population is the group from which you would like to sample and the group to which you are interested in generalizing your findings. For example, you may be interested in staff development for all elementary school teachers (your study population), and plan to draw a sample from this group of teachers.

Once you have identified the relevant population, you have to do one more thing before you can actually draw a sample: You have to get a complete list of all of the members of your target or study population. This listing of the accessible population from which you will draw your sample is called the **sampling frame**.

***The sampling frame is a list of potential study members. It can include individuals (teachers) or organizations (schools).***

If you were doing a telephone survey, for example, you might use a telephone book or list as a way to select individuals to call—in this case, the book would be your sampling frame. But this may not be a good approach for some studies because some people may not be listed or may not have telephones. (For example, if you were conducting a study of poor families, you would miss too many if you depended on a telephone book.<sup>3</sup>)

***The sample is the collection of individuals or organizations that will participate in your evaluation study.***

After developing your sampling frame or list, you are ready to actually draw your evaluation **sample**; that is, the group of people you *select* to be in your study. We will have more to say later about how you actually select the sample, but for now it is important to understand that your actual **study sample** may be different than the sample you select at the start of your evaluation. This difference can arise because of a variety of factors, such as noncooperation (some teachers refusing to complete your survey, or test information not being available for some students in the district's data records) or an inability to locate your sample members within the time available. The group that actually completes your study is a subsample of the initially selected sample; that is, it excludes nonrespondents and, maybe, program dropouts you cannot locate. To the extent that these nonrespondents are different in important ways from those who cooperate with your data collection, your study results will be biased. So it is always important to do your best to minimize conditions such as nonresponse or study dropout.

***How Large Should My Sample Be?*** Because they don't include everyone in whom you are interested, samples have an error associated with making statements about the general population. This is called the **sampling error**. The lower the sampling error, the more reliable the estimates, or statements, you can make about your study population. When developing samples, you will want to consider the following:

- How much do you already know about the population being studied? The less information available, the larger the required sample.
- What is the size of the overall population? The larger the population, the larger the required sample—up to a certain point.
- How variable are the phenomena being studied? The more variable the outcomes of interest, the larger the required sample.

---

<sup>3</sup> One solution to this problem is to identify the area code and all three-digit prefixes within that area code and draw a sample simply by randomly dialing numbers (this approach is known as *random-digit dialing*). In this case, the sampling frame is not a list per se, but is rather a procedure that you follow as the actual basis for sampling.

- How important is the decision that will be made from the study results? The greater the degree of confidence one needs in the evaluation findings, the larger the required sample.
- How reliable is the measure being used? The more reliable the measure, the smaller the required sample.

Computing sample sizes is a common statistical practice, but it is too technical to go into in this handbook. You can consult one of the sources listed in the reference section of this handbook if you would like more information about how to do compute sample sizes, or you can depend on a member of your evaluation team who has had the necessary statistical training.

***How Do I Select a Probability Sample?*** The simplest form of probability sampling is called ***simple random sampling***. An example of this type of sampling would be if you had a list of all your teachers (say there were 500 teachers on the list), and you wanted to select a sample of 50 teachers for your program evaluation. You could put everyone's name into a hat and draw a sample of 50 names. An easier way to do this is to pick a random starting place, and then select every tenth name on the list. This procedure would give you a sample of 50 teachers with the same probability of being selected (equal to 1 in 10) as in the simple random sample. This type of sampling, by the way, is called ***systematic random sampling*** because it is based on a systematic rule for drawing the sample while maintaining the requirement that each individual be selected at random and with a known probability.

Another technique, called ***stratified random sampling***, involves dividing the eligible population into different groups (or strata) that are more homogeneous than the entire population, and selecting random samples within each subgroup. For example, you could divide your teachers by grade level or by their level of teaching experience (or by any other characteristic of importance to you) and then select separate samples of teachers from each group. This helps to ensure that your study group includes representatives of all different types of teachers. More complex procedures use multiple stratification rules.

Similarly, you may want to conduct ***cluster sampling***, which is useful when you cannot create a list of all potential study participants in advance but you do know where they are located (e.g., in particular schools). For example, you could first select "clusters" of teachers by sampling individual schools and then selecting a sample of teachers from within those schools, or you might want to divide your district into "regions" and then draw

a sample of teachers from within each region. In many cases, researchers will combine stratification and clustering both for convenience in drawing the sample and to improve the reliability of the estimates.

A final method of sampling, termed ***multi-stage sampling***, is helpful when it is difficult to enumerate or list all the potential sample members. For example, let's say you wanted to select students for a study but found it hard to create a complete list of all the students in your district. In a multi-stage sample, you could stratify your schools (i.e., student clusters) by grade level (e.g., elementary, middle, secondary), then select a sample of schools from each stratum of clusters (the first stage of selection), and then select students within the sampled schools (the second stage of sampling).

Your choice of a particular procedure will depend upon what you are evaluating, the practical constraints on enumerating all eligible participants, and considerations of sampling error (i.e., stratification improves reliability for the same overall sample size, and clustering worsens the reliability of the estimates). However, random sampling is not always possible, especially in programs involving children and services. You will then want to consider non-random sampling.

### Non-Random Sampling

With probability sampling we know the chance that any one study member would be selected, and can estimate the degree of reliability of any estimates that are derived from the sample (i.e., we can quantify our confidence in the estimates). But sometimes we do not want to use a random sample. Although we would be unable to generalize from the sample (or to determine the reliability of the resulting study conclusions), non-random samples allow evaluators to use their knowledge about the study population to select particular types of respondents that they would like to include in the study. For example, you may know that a particular school or group of teachers would be an excellent choice for your initial efforts at creating a new professional development program related to the new state assessments and would like to focus your initial evaluation on this "demonstration" project.

Non-random samples can be of two general types:

- **Convenience samples.** This method involves selecting a sample based on ease of recruiting participants for the study; examples include the “person on the street” interview.
- **Purposive samples.** This method is used when there is more of a “purpose” in mind; for example, researchers want to get five whites, three African Americans, and two other ethnic group members in our sample of 10 teachers. These proportions may or may not match how the different groups are actually represented in your district.

The actual selection of non-random samples can be done in a variety of ways, including **expert sampling** (a panel of experts make the selections; e.g., principals nominate teachers), **quota or proportional sampling** (we want 50 percent elementary school teachers so we just pick teachers until we reach the desired quota), and **snowball sampling** (we identify a teacher who meets some criteria for entry into the program and ask him/her to identify other teachers who are similar).

Do I Have to Be an Expert?

Sampling is one of the more complex aspects of research studies, and this handbook is not intended to make you an expert. What we do hope, however, is that you can become familiar enough with the concepts and underlying ideas that you can be better equipped to work with a statistician to design your evaluation. As mentioned, the references section of this handbook provides some good reference materials on sampling.

## Making the Choices: Kramer School District Revisited

Returning to our example of the Kramer School District from chapter 6, recall that the staff development office wanted to

- Select 20 elementary schools, out of a total of 60, that would have a mentoring program.
- These 20 schools include about 160 3rd- and 4th-grade teachers. Twenty teachers—1 per school—would be selected to serve as master teachers, and a total of about 60 teachers (an average of 3 per school) would serve as mentees in this program component. (This represents a total of 80 teachers out of the 160 in the 20 selected schools).
- Half of the 80 mentoring program participants (i.e., 40 teachers) would be invited to participate in the summer institute (i.e., the other half would receive only the mentoring program).

- The remaining 40 teachers for the summer institute would be selected as follows: Half (20) would come from the mentoring schools from among teachers who were not participating in the mentoring program, and half would come from the 20 schools that would be “matched” to the mentoring schools.
- Teachers in the 40 schools that were **not** involved in either training component would serve as the comparison group for the evaluation. The 20 matched schools also serve as the comparison group for the 20 mentoring program schools.

We have, therefore, two kinds of samples—a sample of schools and a sample of teachers within schools. This is also a multi-stage sample, since schools are selected first and then teachers (and their students) are selected at the second stage. It is also a cluster sample of teachers and students, as “clusters” (i.e., schools) are selected first and then the teachers/students are selected from within the sampled schools.

The first step that the evaluation team implemented was to collect information on all of their elementary schools, including student and teacher characteristics and prior student performance on the state assessments. The staff then used this information to “match” schools on as many of these characteristics as possible. Although there are statistical ways to do this, given the time and resources they had available, they just did the best they could to come up with 20 study schools, each of which was matched to a similar comparison school (for a total of 40 schools). Because they did not want to make it appear that favoritism was at work, they simply randomly assigned one of each of the school pairs to get the mentoring program, leaving the other to serve as a comparison school. (Using this random process to assign the “treatment” actually strengthened their evaluation design.)

Next, they wanted to select their sample of mentors in each of the 20 schools that would get the mentoring program. After some discussion, they realized that this should not be a probability sample. Deciding who should be a master teacher had to be based on professional judgment, so they turned to the school principals to make this decision. At the same time, they asked each principal from the 40 schools to send in a roster of all of their 3rd- and 4th-grade teachers.

Once they had received all of the school teacher lists and identified the master teachers in the 20 mentoring schools, they next needed to sample teachers to participate in only the mentoring program, only the summer institute, both programs, or neither program. They decided to do this using systematic random sampling from the teacher lists (as described earlier in this chapter) to yield the final samples shown in the chart on the next page.

**Exhibit 6: Example of Kramer School and Teacher Sample with Multiple Treatments**

<b>Study Group</b>	<b>Mentoring Schools (N = 20)</b>	<b>Comparison Schools (N = 20)</b>
Mentors	20	0
Mentor only	10	
Mentor + summer	10	
Mentees	60	0
Mentee only	30	
Mentee + summer	30	
Summer institute	20	20
Neither program	60	140
Total 3rd- and 4th-grade teachers	160	160

## Chapter 8. Data Collection: Choosing the Methods

### Deciding How to Collect the Data

Once you have decided what information you need to collect and identified your relevant study population and sample, the next step is to determine *how* to collect your data. To some extent, your selected research design(s) will favor particular methods of data collection over others, but a good place to begin thinking about collecting data is to ask, “What data sources will be used?”

Types of data sources that you are likely to encounter:

- **Existing information.** You don't always have to go out and collect new information. You should always explore ways to use readily available information such as school or district records (including student test scores), survey results from the previous year, lesson plans, and other available information.
- **People.** These include program participants and program staff (e.g., administrators and trainers).
- **Observations.** Direct observation of program activities is an underused but powerful tool, and can include efforts to document what happened, having observers rate the “quality” of the activities and settings, and having observers assess intermediate or final program outcomes (e.g., changes in how teachers behave in their classrooms).

Of course, you do not have to select a single source; as with research designs, multiple sources are almost always better. As mentioned in chapter 5, three types of effects can be measured: effects on teachers, effects on students, and effects on the organization. Below are examples of typical data sources for each of these effects:

- **Effects on teachers** (or other staff) can be measured using participant surveys, supervisors' assessments, interviews, portfolios, and classroom observations.
- **Effects on students** can be measured using student surveys or interviews; standardized tests or authentic assessments and portfolios; grades; information on attendance and tardiness; dropout and retention rates; and information on disciplinary actions or school vandalism.
- **Effects on organizations** can be assessed by reviewing the minutes of meetings, formal changes in policies and procedures, changes in the allocation of resources, and the creation of new governance structures.

Gathering data from a variety of sources, and using a range of methods, will strengthen the validity of evaluation findings and allow you to **triangulate** your results. As mentioned in chapter 6, using multiple respondents and methods of data collection can allow you to confirm your results through various sources. For example, teachers may self-report that their teaching is well aligned with the state standards. These findings can be confirmed through classroom observations and interviews with students, among other means. Sometimes triangulation yields a deeper understanding of how well a program has been implemented and may shed light on your findings regarding its effectiveness.

To decide which data collection methods are most appropriate for your evaluation, you will want to consider the following factors:

- **Which method is most appropriate given the source and/or information needed?** Some methods are more appropriate for collecting certain types of information (e.g., tests to gauge changes in knowledge or skills). For example, surveys might be best for assessing changes in teacher attitudes and/or beliefs, tests might be best for measuring changes in teacher knowledge or skills, and observations might be best for capturing changes in classroom instructional practices.
- **Which method is least disruptive to your operations?** Administering standardized tests can be quite time-consuming and disruptive of school time. In many cases, surveys or interviews, which can be done “off hours,” may have a far lower impact on school operations.
- **Which method can you afford?** Not all methods are equally costly, so you will have to estimate the cost of different modes of data collection and assess the trade-off between validity and what you can afford. Self-administered teacher surveys are less expensive than classroom observations, but observations can yield better information about what teachers are actually doing.

You should select methods that best meet the objectives and goals of your evaluation and that will appear most credible to your stakeholders. For some questions—and in cases where the decisions to be made from the evaluation are high-cost choices—a randomized experiment combined with the most rigorous methods of data collection will be the only valid research strategy; for other types of questions, less robust designs may be appropriate. Having credible evidence also strengthens the ability to take actions and make recommendations based on the evaluation results. Exhibit 6 provides some information on the more common methods of collecting data and their relative strengths and weaknesses. Each is discussed below.

**Exhibit 7: Common Methods of Data Collection**

<b>Method of Data Collection</b>	<b>Advantages</b>	<b>Disadvantages</b>
<b>Self-administered questionnaires</b>	Inexpensive and easy to implement. Good for short and simple surveys.	No control over who completes the form. No ability to clarify misunderstood terms or questions. Not well suited for complex issues. Self-report may not match actual behavior.
<b>Interviewer-administered telephone questionnaires</b>	Relatively inexpensive, and avoid need to send staff into risky neighborhoods. Also better suited for short and simple surveys.	Telephone coverage is a problem for low-income respondents; also not well suited for children, elderly, and non-English speakers. Not good for complex issues. Self-report may not match actual behavior.
<b>Interviewer-administered in-person questionnaires</b>	Can probe for more in-depth answers, and can explain confusing questions. Personal rapport can increase trust.	Expensive, and often require lengthy data collection period. May present logistical problems related to gaining access to respondent. Self-report may not match actual behavior.
<b>Interviewer-administered, in-person, open-ended interviews</b>	Yields rich in-depth data.	Same as above, but data are also often more difficult to analyze. Self-report may not match actual behavior.
<b>Focus groups</b>	Useful for gathering ideas for later study in a broader survey. Allow new insights to be gained; helpful in questionnaire design.	Not suitable for making generalizations about the population being studied.
<b>Tests</b>	Relatively easy to administer, and can use commercially available products. Provides "hard" data on performance outcomes.	Instruments may be unsuitable for particular population; developing new tests can be very expensive.
<b>Observations</b>	Best way to collect information about behavior of individuals and groups.	Usually very expensive, and require well-qualified staff. Can be difficult to gain reliability across observers. Actual behavior can be assessed directly, but choosing appropriate time samples is difficult, and observers may affect behavior being studied.
<b>Document and record reviews</b>	Avoid need to collect new data. Very inexpensive.	Limited by the availability and quality of existing data systems.

(Adapted from Stevens et al. (1993), p. 44)

**Self-administered questionnaires** were used in the recent 2000 Census. Survey forms were mailed to every household in the United States, to be completed and returned by the head of household. This is a relatively inexpensive method of data collection that can be easily distributed to large numbers of respondents. Drawbacks include the lack of control over who actually completes the form, inability to clarify misunderstood questions or terms, limited ability to explore complex issues, and self-report bias—the tendency for respondents to provide information that does not match their actual behavior or opinions.

A new variant of this method is the use of the Internet for Web-based survey administration. For school personnel, this could involve the use of email or internal network Web sites to distribute the survey and for teachers or other staff to complete the form on line.

The next two categories introduce the use of an interviewer—hence the name **interviewer-administered**—who can interact with the respondent. These interviews are typically conducted with “key informants,” with a written instrument that includes information about the purpose of the interview, the level of confidentiality to be provided to the informant, and other procedural details to guide the interview.

***There are four types of interviewer-administered questionnaires: those conducted over the telephone versus those conducted in person, and those that use primarily structured questionnaires versus more open-ended interviews.***

**Telephone interviews** involve highly structured questionnaires<sup>4</sup> or more loosely structured protocols that are typically short due to the high likelihood of losing the respondent if the interview drags on too long. This form of data collection is relatively inexpensive and, in larger surveys of American households, can be administered using modern computer-assisted systems (called computer-assisted telephone interviews, or CATI) that allow for “random-digit dialing” to sample respondents and for simultaneous data editing and checking. However, telephone interviews are often problematic if respondents do not have easy access to a telephone (e.g., teachers lacking telephones in their classrooms) or for children, certain disabled individuals, and individuals with limited English-speaking ability.

An alternative to the telephone interview is the **in-person interview**, which provides an opportunity to capture more in-depth information from program participants, staff, and

managers, as well as others with a stake in the program. An in-person interview is usually selected when the personal contact is important, such as when follow-up questions can be asked, points clarified, and/or the discussion can lead into unexpected areas or aspects of the program.

Such interviews can be primarily structured or open-ended:

**Structured interviews.** A carefully worded questionnaire is used to obtain the same information from all respondents. Interviewers are trained to deviate only minimally from the questionnaire, and the questions are designed to ensure consistency.

**Unstructured or “open-ended” interviews.** Interviewers do not follow a rigid format but rather use a more conversational style designed to explore key themes. Interviewers seek to encourage free and open responses, and there may be a trade-off between breadth and in-depth exploration of selected topics.

In the first case, the structured interview involves simply administering the questionnaire orally (as with the structured telephone interview). Interview results in a structured interview are most often recorded on the survey form: The interviewer records the responses with marginal notes where necessary.

Alternatively, an open-ended interview guide, often called an interview “protocol,” contains broader questions that allow the interviewer to explore evaluation issues with the informant in greater detail than through a structured questionnaire. In addition to guiding the interview, protocols often include optional questions and/or notes to help the interviewer gain a more specific understanding of those issues that are important to the evaluation. The protocol includes these questions, or “probes,” to alert the interviewer to areas of interest for the evaluation. The probes enhance the power of the protocol to gather rich and useful information.

In an unstructured interview, the interviewer will typically take notes that must later be converted into a “story” of what was learned. Tape recording (with the respondent’s permission) may be helpful to avoid the delay involved in taking copious notes and to ensure that nothing is missed. Thanks to recent technological developments, interviewers can also use laptop computers to aid the interview process and to allow “instant” recording of the data obtained (this is called computer-assisted personal interviewing, or CAPI).

---

<sup>4</sup> Telephone interviews do not have to be structured but are commonly limited in this way due to the time and other constraints imposed by this mode of data collection.

In-person interviews are typically expensive, require an extended period of time to complete (e.g., scheduling the interviews with all the respondents), and require extensive training of interviewers, especially to ensure consistency across respondents.

***The focus group session is an interview involving multiple participants.***

In **focus group interviews**, the interviewer or moderator follows a protocol that helps guide the discussion among multiple respondents. These protocols resemble those used in open-interviews with a single participant, and may include probes that remind the interviewer of key topics to be covered.

The advantage of a focus group is that it allows participants to interact and expand on one another's answers and observations, allowing for a more enriched perspective on the topic. The discussion is led by the moderator, who keeps the discussion on target, ensures that everyone contributes, and focuses the discussion on the key questions to be addressed by the group. Usually 7 to 10 people are interviewed, and the discussion can run for up to about two hours. Taping may be used to record the discussion, or an independent "recorder" can be used to take notes.

Focus groups are most appropriate for identifying problems in program implementation, identifying participant needs, generating new ideas for program improvement, testing the validity of insights about the program, and assisting in the development of data collection instruments for a broader study. They are also useful when you want to find out about the dynamics of how people interact to help deepen your understanding of the program.

***Tests are usually standardized measures that capture an individual's level of knowledge, skill, or performance ability.***

**Tests** and other types of standardized assessments are a common part of the school environment, and their use has grown in recent years with the greater push for increased accountability. Tests are often available from commercial suppliers, are relatively easy to administer to a large group of individuals, and can provide good quantitative measures for later analysis. However, the available tests may be inappropriate for your particular evaluation (e.g., the tests don't capture the learning or changes you expect to see). In fact, a key feature of most standards-based reform movements at the state and district levels is

that the tests must be *aligned* with the standards and the curriculum being taught. For this reason, many states and districts have been devising or revising their student assessments.

In addition, the tests may be inappropriate for your study population (e.g., very young children). Developing new tests for your evaluation is not recommended, as this requires considerable expertise and resources.

***Observations involve the use of trained individuals who can observe actual program activities to better understand the context within which operations occur and how the program operates, and in some cases can be used to “rate” the quality of service delivery.***

**Observations** of participants in their “natural” setting (e.g., teachers and students in a classroom) provide direct information on the behavior of program participants and service activities, provide insight into the context within which the behaviors naturally occur, allow for the identification of unintended consequences, and allow staff to examine outcomes in the natural program setting.

Of particular interest for professional development research is that classroom observations provide an opportunity for observers to assess how professional development training is being put into practice. Someone familiar with a professional development program may be able to rate teachers’ level of implementation of a teaching strategy or approach in the classroom. A skilled and trained observer may even determine students’ responses to the instruction that is delivered by teachers who participated in professional development.

On the other hand, observations are expensive and time-consuming to conduct and require well-trained staff to do the observations. The observer may also affect the behavior of the participants, observations may be affected by the observer’s selective perception, the investigator may have little control over the situation, and the behaviors observed may not be typical.

Observations are typically guided by protocols, and, like interview protocols, observation protocols can be more or less structured. Structured protocols typically require the observer to make judgments about various behaviors using checklists, rating scales, and other specific data points. For example, an observation protocol may ask the observer to record the number of times professional development participants appear to be focused on a specific activity, to keep track of the number of minutes spent in each type of activity

(e.g., collaborative work with other participants, listening to the presenter, asking questions, taking a break). The unstructured protocol requires that the observer record his or her impressions of the activities, typically with some guidance on the protocol about issues to be observed.

***It is particularly important, when conducting observations or other using qualitative methods, to train the data collectors on the protocol or questionnaire that is to be used in the field. Even seasoned evaluators need to sit down and discuss the specific language of the instruments to ensure a consensus. Without this training, the use of qualitative measures can be inconsistent, compromising the quality of the data.***

The final method of data collection, ***document and record reviews***, capitalizes on the availability of existing data that may be relevant to your evaluation. Existing records can often provide an important insight into the program. These can include historical accounts, mission statements, plans, annual reports, budgets, test-score data, lesson plans, minutes of meetings, internal memoranda, correspondence, policy manuals, mass media reports, and so on. Such sources often help provide insight into the program context and planned operations, and in some cases (e.g., test scores, attendance records) can provide a way of tracking performance outcomes over time. This first type of data collection does not require intrusion into the day-to-day operations of a program.

- ***Indirect measures.*** These are measures that occur naturally in the program environment. For example, in a training workshop with “break-out” sessions that focused on different topics that might be of interest to your teachers, keeping track of the number of teachers who attend the different sessions would give you some indication of what the staff see as the most important topics.
- ***Content analysis*** This is the analysis of text documents such as curriculum guides and lesson plans. The analysis can consist of looking for key or repeating themes or keeping an index of the number of times a certain word or phrase appears.
- ***Secondary analysis*** This final method involves the analysis of existing quantitative data, such as an analysis of student test scores, incidence of absenteeism or disciplinary actions, or other such data that are already maintained in your organization.

## Data Collection Methods for Kramer School District

Let's return to our friends in the Kramer School District, where the staff development team working on the evaluation realized that they had to collect the following types of information:

- **Effects on teachers**—(a) changes in knowledge, attitudes, motivation, and self-efficacy, and (b) changes in classroom practices.
- **Effects on students**—(a) changes in classroom behavior/engagement, and (b) changes in academic achievement.

The first group of teacher effects data would be collected through self-administered questionnaires, the second through the use of interviews with school administrators and classroom observations. Student effects would be measured using existing student achievement test score data and as part of the classroom observations planned for teachers.



## Chapter 9. Data Collection: Creating and Using the Tools

### Constructing Data Collection Instruments

School and district staff do not generally have the time to develop and test questionnaires or other types of data collection protocols (e.g., observation guides, interview protocols, focus group guides, etc.) for use in evaluation studies. To the extent possible, therefore, staff should attempt to locate existing—and tested—data collection instruments that meet their needs. In some cases, just a portion of a longer questionnaire (even single questions) is relevant, and it is acceptable to just use the parts that are appropriate for your evaluation. However, questions from widely used surveys should *not* be changed if you intend to make comparisons to prior research results as part of your analysis and interpretation of your evaluation findings.

If you cannot find an appropriate instrument, new questionnaires or other tools must be developed. Some useful guidelines for doing this include the following.

#### Determine the Content and Purpose of Your Questions

***You will want to be sure that every item in your data collection instruments is linked to a research question and measure.***

To ensure that this is the case, use the matrix discussed in chapter 7 to link your questions to a particular research purpose. In many instances a single question will not be sufficient to capture all the data you need for a particular research question or measure. For example, assessing a teacher's self-perceived efficacy may require the use of several items that are later combined into a single "scale," a collection of variables that together tap into a particular psychological construct. You also may want to first determine whether something occurs (e.g., "Do you do X in your classroom?"), and the frequency with which it occurs (e.g., "How often do you do X in an average week?"). It is also often the case that some questions will support more than one research question or measure. Be sure that every question you ask has a clear purpose. Every question imposes a burden on your respondent.

Other things to consider in developing the content of your instruments:

- **Does the respondent have the information?** Avoid asking questions that the individual cannot answer.
- **How specific should the question be?** Too often, survey questions are too general to be useful. For example, if you want to learn about teachers' opinions of the training program, you could ask, "How well did you like the training?" But you would get a much better understanding if you posed several questions that targeted specific components of the training and their perceived utility.
- **Is the question sufficiently general?** It is also possible to be too specific. For example, you might be interested in the extent to which teachers use particular teaching strategies in their classroom. One question might ask, "Did you do X in your class last week?" A "no" to this question is clear, but it might also mean that the reference time period is too narrow, especially for events that are more episodic (e.g., occurring once or twice per month).
- **Is the question biased?** "Loaded" questions, where the question wording tips off the respondent as to the answer being sought, are a frequent problem in surveys. As a consequence, be sure to keep wording as neutral as possible.
- **Will the respondent answer truthfully?** The final question to ask yourself is whether there is any reason for the respondent to provide intentionally incorrect answers or to feel he/she should answer in a particular (socially desirable) way. This can be a problem when you are asking potentially sensitive questions (e.g., those dealing with substance abuse) or questions for which the respondent may perceive some risk to their job status (e.g., the use of classroom practices that run counter to "official" policy).

### Consider Different Types of Measurement Tools

Different methods of data collection require different types of tools to gather the information from your evaluation sample, and in some cases you can use different types of tools to measure the same thing. For example, if you wanted to measure the distance between two points, you could use a yardstick, a tape measure, precision calipers, or even a sophisticated laser device. Each would give you an answer, but the accuracy of your measurements would vary, as would the cost of collecting the data. In part, the choice of the best tool would depend upon the risk associated with getting the wrong answer. In evaluation, the problem is the same: We want accurate answers, but how we go about taking our measurements must reflect both the importance of being accurate and the available resources.

By **measurement** we mean the process of observing and recording the information that will be collected as part of your evaluation. Two major issues need to be considered in deciding how to take those measurements: the **level** of the measures to be used, and their **validity** and **reliability**.

## Understanding Different Levels of Measurement

- There are many different types of measures, or what researchers call “variables,” that can be distinguished on the basis of the values assigned to the different “levels” or categories that are being measured:

- **Nominal**—The weakest level, in which the numerical values just “name” the attribute, with no implied ordering, such as jersey numbers assigned to members of a baseball team (i.e., number 30 is not twice as good as number 15).
- **Ordinal**—Attributes can be rank-ordered, but there is no particular meaning to the distance between adjacent ranks, such as the coding of highest level of educational attainment as 1 = less than high school, 2 = high school, 3 = some college, 4 = two-year college degree, 5 = four-year college degree, and 6 = Masters degree or above. The categories represent progress along an educational scale, but the jump from 2 to 3 is not equal to the jump from 5 to 6.
- **Interval**—Here the distance between different values has meaning, such as a temperature scale where, for example, the distance from 30<sup>o</sup> to 40<sup>o</sup>F is the same as the change from 70<sup>o</sup> to 80<sup>o</sup>F. In other words, the interval between measurements can be interpreted.
- **Ratio**—This category includes cases where there is a defined meaning for the value of zero and you can construct a meaningful fraction (or ratio)—for example, a count of the number of teachers who received training.

Understanding these different types of measures is important because it affects what you can do with the data in terms of analysis. In general, having a higher “level” of measurement (e.g., interval or ratio) is preferred, as it allows you to do more analysis of your data.

## Measurement Validity and Reliability

Two other important concepts for developing measures are the validity and reliability of your measures:

- **Validity**—The information (e.g., a particular questionnaire item) must measure what it actually claims to measure. For example, if you are interested in the self-perceived efficacy of teachers, you will want to select measures of this concept that, to the extent possible, are valid representations of how teachers believe they have the ability to effect change in their students.
- **Reliability**—In addition to being valid, measurements made of the same situation should yield the same results; that is, the measures should be stable across unchanging situations. This is really about consistency—if I ask the respondent a

question today, tomorrow I should get the same answer, as long as conditions have not changed.

Can a measure be reliable and yet not valid? Yes. For example, if you ask students about some illicit behavior this week and next week, and each time ask them to sign their name to the questionnaire, you may obtain highly reliable (i.e., consistent) information but it may be totally wrong if students are not answering correctly out of fear of retribution from school officials.

#### Choose the Appropriate Response Format

As noted above, two types of question format exist. The first, an **unstructured question**, does not offer the respondent a set of possible responses; an example would be, "Tell me about how you currently teach math." The response to such a question is usually recorded verbatim for later analysis. More experienced researchers can record just the gist of the respondent's reply to these types of question.

The second format is a **structured** question. Several types of structured question can be used, each with a different approach to how response possibilities are provided to the respondent. Some of the more common designs are shown on the next page.

#### Determine the Best Wording

Writing good questions is a difficult task; even slight wording changes can affect how the respondent interprets the question and the quality of the data you obtain. When preparing questions, consider the following:

- *Does the question contain difficult or unclear terminology or language?*
- *Does the question make the response choices clear?*
- *Is the wording objectionable or "loaded"?*
- *Can the question be misunderstood?*
- *What assumptions does the question make?*
- *Is the time frame specified?*

### Types of Structured Questions

The simplest type is the “fill in the blank”; for example,  
*How many years have you been teaching? \_\_\_\_\_*

The next type is the *dichotomous choice*; for example,  
*Please indicate your gender*  
Male \_\_\_\_\_  
Female \_\_\_\_\_

You can also use questions that depend on the level of measurement. This can, for example, include nominal level (simple naming) questions:

*What grade do you currently teach?*  
1<sup>st</sup> grade \_\_\_\_\_  
2<sup>nd</sup> grade \_\_\_\_\_  
3<sup>rd</sup> grade \_\_\_\_\_

Or, for example, ordinal (ranking) or interval-level (leveled ranking) questions:

**Ordinal:** *Please rank, in order from most to least important, the topics that you would most like to see included in professional development activities:*

Item 1  
Item 2  
Item 3  
Etc.

**Interval:** *“Please indicate how important each of the following topics are to your continued development as a classroom teacher:*

	Very Important	Somewhat Important	Somewhat Unimportant	Very Unimportant
Item 1				
Item 2				
Item 3				
Etc.				

You may also find the use of *filtering* questions useful as they allow your respondents to answer only those questions that apply to them:

*Do you use computers in your classroom?*  
No \_\_\_\_\_  
Yes \_\_\_\_\_  
  
*If yes, how many do you have? \_\_\_\_\_*

When using such filters, be careful to avoid including too many “jumps” for any one question, as this may confuse the respondent.

Here are some examples of questions that could be difficult for respondents to interpret:

- *Is your school's technology plan aligned with the state standards?* Without a definition for “plan”—such as “a formal written plan for the acquisition and use of instructional technology”—respondents may interpret this to mean something more informal than is intended, such as district staff’s loose sense of where the school is headed regarding technology.
- *What are the goals of your school’s educational technology plan?*
  - a. *To provide professional development for teachers on the use of technology* \_\_\_\_
  - b. *To provide professional development for improving academic instruction* \_\_\_\_
  - c. *To provide technical support for teachers* \_\_\_\_
  - d. *To make modern computers available in the classroom* \_\_\_\_

Can respondents check all that apply, or should they restrict their responses to the most relevant? Even inserting the word “major” before goals does not ensure that respondents will use the same criteria when answering the question.

- *To a parent: Are you familiar with the state standards?* Several confusions could result. Does the question mean “you” as in only yourself, or does it include other parents and members of the household? Does “familiar with” mean having a general idea of what the standards are, being familiar with the content, or having studied the standards? Even “state standards” is confusing: Does the question refer to one subject, core subjects, or all subjects? The time period is also unclear. What if the respondent reviewed an earlier version of the standards several years ago? One way to clarify this question without resorting to providing several definitions is to ask a series of questions, such as, “Do you know that state standards have been developed in the following subjects? Have you read them? Have you discussed them with your child?” Of course, crafting these questions will depend entirely on your information needs.
- *In the classroom, are you a facilitator of student learning?* Given school reform’s current emphasis on teachers having a more facilitative role in the classroom (the “guide on the side” instead of the “sage on the stage”), this question could be considered biased. Many respondents will know that they are “supposed” to take on this role, and may have a hard time acknowledging (or even seeing) their deficiencies in this regard. A better question would focus on specific behaviors, such as whether and how often teachers ask their students to help develop assignments or work on independent projects.

Other issues to be careful about include the following:

- (1) The use of vague qualifiers such as “never,” “rarely,” and “often.” Respondents will interpret these qualifiers very differently. In fact, distinct gender differences—and lesser differences along race and socioeconomic lines—exist in the interpretation and use of these terms. Using open-ended questions or numerical scales (e.g., a scale of 1 to 4, where 1 means never and 4 means always) avoids these ambiguities.
- (2) Presuppositions embedded in your question, such as, “How many technical difficulties have you had with educational technology?” The questionnaire developer has

assumed that respondents have had at least one. Even if you include a “none” category, the question tends to bias the respondent.

### Decide on Question Placement

The final task in developing questionnaires is deciding on the order in which the questions will be asked. This is not an easy task, but here are some useful guidelines:

- Keep questions on the same topic together. Avoid sharp “jumps,” and be sure to use transitions to indicate a change in topic (e.g., “Below are some questions about your home and neighborhood”).
- Start with easy, nonthreatening questions. Put more difficult or sensitive questions near the end.
- When asking sensitive questions, include language that acknowledges this and, if necessary, indicate that the respondent may choose not to answer selected questions. (At the same time, do not make it easy for respondents to skip questions they would prefer not to answer.)
- Be careful to not place all of your most important questions at the end, where respondent fatigue may reduce the accuracy or completeness of the data.

### Pre-Testing Your Data Collection Instruments

***It is important to pretest any questionnaire or protocol using the same procedures that will be used in the actual evaluation, and with similar respondents—though not with the actual evaluation subjects.***

The pretest will allow you to identify any problems that need to be fixed before full-scale data collection. In a pretest, a mock respondent completes the survey or interview and discusses the instrument’s relevance, clarity, and suitability for use in the field with the instrument developer. This “debriefing” of the respondent also assesses whether the questions were correctly interpreted and whether there were any confusions or problems with specific questions or the overall survey or protocol format. Respondents may also be asked to share their impressions of how the instruments could be strengthened.

Typically, evaluators seek to pretest instruments with as many people as possible who resemble the actual subjects for the evaluation—but not, unless absolutely necessary, with individuals who will be part of the evaluation (because they may learn, through the debriefing, about the *intentions* of the evaluation team and then become less objective as an informant).

## Managing Your Data Collection

A variety of additional tools can make your data collection more efficient. First, a concise **study description**, written in simple language, is a helpful product that can be used to inform potential study participants about the evaluation, as well as to brief stakeholders about your plans. It can be a simple one-page summary or a more elaborate study brochure. The key is that the study description should help answer most (if not all) of the questions people will have about your evaluation and help them understand the study's importance.

Next, if you are planning to send out self-administered questionnaires, you will want to prepare **letters and instructions** that inform the recipient of the purpose of the study and provide instructions for completing and returning the information to you. In some cases, additional letters of endorsement from third parties (e.g., union representatives) can help gain needed cooperation.

Finally, you certainly will want to create a **system for tracking** the status of your data collection effort. This is important regardless of the type of data collection you are doing—surveys, observations, or interviews. In most cases you will want to create an electronic database for this purpose. You can use a variety of software packages for this purpose, including spreadsheets (Excel, Lotus), databases (Access), and analysis tools (SAS, SPSS). Each study participant should be given a unique identification number for tracking purposes (you can have a separate code, for example, to designate the school and a code number for a particular teacher within the school), and you will want to keep track of the status of each data collection form for that individual case (e.g., not done yet, in process, completed, data being key entered, etc.).

The important thing to keep in mind is that you will want to be able to monitor the status of your data collection at all times. This will allow you to determine how many surveys, interviews, and so on you planned; how many were completed (i.e., your response rate); and where you failed to obtain data. Keeping track of data collection will also help you link data back to the original data collection form and will be important for follow-up evaluation activities. For all but the most rudimentary data collection efforts, this type of tracking system is a must; without it you will not be able to know where you are nor when to eventually cut off further data collection. You also won't be able to report your eventual response rate.

## Logistical Planning for Data Collection

Regardless of the method of data collection that you decide to use, you should incorporate some common logistical considerations into your plan, as they have implications for both the quality and cost of your evaluation. Examples include the following:

- **How often will the data be collected?** Should it be collected once, at multiple time points, or continuously? If at multiple times, when? Obviously, this will depend at least in part on your selected research design (e.g., are you planning a pre-test only? pre-test/post-test? multiple pre- and/or post-tests?).
- **Who will collect the data?** Can you do it yourself, or do you need outside help?
- **When will data be collected?** When will information be available? When can it be conveniently collected? Where will the information collection take place? When will data collection start and end? Consider your respondents; convenient times will differ if you are collecting information from teachers, students, administrators, or parents.

In addition to these issues, you will also have to consider how you will recruit/select, train, and supervise your data collectors if you decide to do this yourself. Interviewers are a critical cog in the data collection machine, especially if you are planning to use unstructured interviews, observations, and structured tests of teacher skills that require higher-level skills and experience. In many cases, interviewers are required to gain the cooperation of the respondents, motivate them to do a good and thorough job of providing information, clarify any confusing items in the survey, probe for more in-depth information, and assess the quality of the reported data. These all require skill and training.

### Selecting Data Collectors for Interviews and Focus Groups

Staff collecting the data will need to have sufficient time to devote to data collection and should be comfortable handling difficult and changing situations. Flexibility is key! You will also want staff who are self-motivated, can pay close attention to details, have strong interpersonal and communication skills, and can work with limited supervision. Prior experience with similar work would be an enormous advantage, of course. Other skills that may be important for your particular evaluation include knowledge about the topic under investigation, so that, for example, they know when and how to probe respondents for more information. Language fluency may be important if you plan to interview respondents who do not speak English. In some situations, using ethnically matched interviewers may be important. When collecting data about race issues, the race of the interviewer has been shown to affect the responses.

## Training Data Collectors

Even if you use experienced data collectors, you will want to provide training for all your staff. Training puts everyone “on the same page” regarding the goals and objectives of the study and ensures consistency of data collection across multiple individuals.

Some of the major topics that should be included in a training session include the following:

- **Study description.** Data collectors need to be able to describe the purpose of the study to respondents, and need to understand the overall study objectives so that they are better able to collect the information needed.
- **Background on survey research.** This is not meant to be a course on data collection, but you should cover good practices and the underlying rationale for your data collection plans.
- **Study methodology.** Explain how the study was designed, how participants are being selected, and how the data are being collected and analyzed. Also include a schedule of major milestones, especially of when you plan to issue a final report on the results of your evaluation and to whom it will be disseminated.
- **Review the questionnaire.** Go through each question and explain what is intended, clarify any confusing issues, describe the types of responses you are expecting, and discuss probes where relevant.
- **Role plays.** A good way to give staff some practice (and to allow you to assess their skills) is to conduct a variety of mock interviews. Staff can alternate being the interviewer and the respondents and can role-play particularly difficult respondents or situations.
- **Interview materials.** Discuss procedures for reviewing and completing all data collection instruments and how they should be submitted to you for processing.
- **Scheduling and supervision.** Review the expected work schedules and how you plan to supervise their time and activities.
- **Administrative details.** Keep track of time sheets, expense reports, and payment plans.

A very useful component of any training session is the preparation of an “interviewer manual” that includes all the information covered during the training session. This can serve as an excellent refresher training source and a handy source of information during the course of the ongoing study. This kit—typically a three-ring binder—should also include master copies of all data collection and administrative forms.

## Nonrespondents

Any data collection effort will have individuals from whom data were not collected. You will want to know the extent of this problem, as it can affect the reliability of your results. In particular, it is important to determine whether you have systematically failed to obtain data from certain types of respondents. If so, this may indicate that your data are biased. This is why it is important to do the best you can at getting responses from a high percentage of your original sample—generally acceptable standards are at least a 75 percent response rate.

Evaluators use several methods to try to get nonrespondents to provide data. These include sending multiple versions of survey instruments with reminders, sending follow-up postcards, making reminder phone calls, and attempting to reschedule other data collection activities, such as interviews and focus groups. Used wisely, these methods can increase your follow-up rate significantly.

If resources and time permit, you may also want to do a follow-up survey of a small subsample (e.g., 10 percent) of the nonrespondents to learn why they did not respond and whether their characteristics are unique in some way. This can help you determine if there is any response bias.

## A Note about Using Data from Existing Records

If you are using data from existing records (e.g., school records of attendance, disciplinary actions, and test scores), it is important to learn as much as you can about the quality of the data before you collect and use it. Issues to be considered include the scope of available data, how the data are defined, procedures used to collect and check the quality of the information, the completeness of the data records, the accuracy of the data, and the timeliness of the information. In many cases, it is useful to combine existing data with information collected from surveys or other forms of primary data collection, as this will allow some checking for consistency across different data sources.

## Protecting Human Subjects and Maintaining Confidentiality

Collecting information from individuals can be a sensitive issue, especially when the data collection involves children. As a consequence, it is always important to inform all study participants about the purpose of the study, how the data will be used, and how their confidentiality will be protected. In addition, in many instances formal written consent

should be obtained. Rules governing these issues are often spelled out in existing state, district, or school policies.

Sometimes evaluations collect personal information about people. If not properly handled, this information can become known to individuals who can do some harm to the respondents, or it can be embarrassing to have the personal information made public. The fear of this exposure may also increase respondent resistance to participation in the study or may result in distorted or biased answers. The best solution to this problem is to pledge **anonymity** to every respondent. A statement such as the following can be used on questionnaires:

***Sample Confidentiality Statement:***

All responses to this survey will be strictly confidential. You will never be identified by name or in any other manner that will allow another researcher, government official, or member of the public to infer your identity.

You need to ensure that such confidentiality is scrupulously maintained throughout the project. For example, individual identifiers should be stripped from all records, data should be kept in a safe place, and data should never be reported in a way that would permit someone to identify a particular individual.

## Data Collection in the Kramer School District

### Instrument Development

The goal of the Kramer School District evaluation was to capture development's effects on teachers and students. The team determined that four types of data collection instrument had to be developed:

- A teacher questionnaire;
- A protocol for the principal interview;
- A classroom observation instrument; and
- A student record abstraction form.

The evaluation was fortunate to have the assistance of Dr. Frank, an expert in educational evaluation from the local university. He was able to locate several teacher surveys that had been used in previous studies of teacher professional development, and with a

relatively small amount of work was able to create a teacher survey that captured the information Kramer School District needed. The survey included several reliable scales of teacher knowledge, attitudes, and self-perceived efficacy that were then combined with questions that were specifically targeted to the respondent's knowledge and practices regarding the state standards and assessments. Similarly, Dr. Frank was able to find several classroom observation instruments that had been developed and used by other researchers and that were easily adapted to the needs of the Kramer School District's evaluation.

The principal interview was planned as an open-ended discussion that would focus on effects observed for the participating teachers, as well as any organizational changes that might have occurred as a consequence of the training program (e.g., improved school climate, increased teacher collaboration). The evaluation team was able to create this interview protocol themselves based on both their knowledge of what information they wanted to obtain and their experience with the schools and administrators in the study.

The collection of student data from administrative records required the assistance of staff from the district's Office of Information Technology. Because the district had a well-developed information system, this task was relatively easy. Once the teachers who would participate in the evaluation were sampled, the evaluation team was able to obtain lists of students assigned to each teacher for (a) the year before the training was implemented, (b) the year in which the training occurred, and (c) the year after the training was completed. These student lists were then matched to the computer files maintained by the Office of Information Technology, to produce an electronic file that linked student-level data to individual teachers.

### Pre-Testing

To ensure that the teacher survey, the principal interview, and the classroom observation instruments would work as expected, the evaluation team tested them out in several elementary schools in the district that were **not** included in the study sample. Based on this pre-test, several questions had to be re-worded and some additional instructions were added to make sure that the respondents were clear about the intent of each question. The final versions were sent to the printer to make sufficient copies for the evaluation.

## Choosing Data Collection Procedures

To implement the planned data collection, the evaluation team came up with the following strategy:

- Members of the staff development office would be responsible for distributing the teacher questionnaires to the selected study teachers' school mailboxes. The same staff would also handle the principal interviews.
- The classroom observations would be conducted by graduate students from the local university. The team considered using district staff for this activity, but a concern for complete objectivity led them to choose external observers. Dr. Frank would be responsible for recruiting, training, and monitoring the students.

The Office of Information Technology would arrange to have all of the teacher surveys and classroom observations key-entered to create an electronic data file for later analysis. They were also able to link these data with the information obtained from existing administrative records.

## Staff Training

The classroom observations required that the data collectors be able to achieve a relatively high level of consistency across observers. This "inter-rater" reliability ensured that there would be only very small differences across teachers that could be attributed to the use of multiple individuals to collect the observation data. Dr. Frank developed a rigorous training regimen for the graduate students, and used videotapes to test the consistency of ratings across the different individuals. Only when the graduate students could demonstrate the desired level of reliability were they deemed ready to conduct the classroom observations.

In addition, graduate students were trained to conduct the principal interviews using the protocol. Every question and probe on the protocol was explained to the students so that they knew what the questions meant and could provide clarification to the interviewee when requested.

## Chapter 10. Data Analysis: Understanding What Happened

**Y**ou and your students have returned from your trip with loads of collected information. Now it's time to figure out what you learned. The next step, then, is to assemble and analyze your data.

### Getting Your Data Ready for Analysis

You are eager to get into the data, but before you start it pays to spend some time getting everything ready so that you can focus on understanding the information that you have collected.

***Every evaluation should begin with a plan for how the data will be analyzed. This plan should link research questions to the specific data being collected, and spell out in detail how the data will be analyzed to answer the research questions.***

### Analysis Plans

Analysis plans should be prepared *at the beginning of the project* and *not* after the data have been collected. An analysis plan should be the roadmap that will ensure that all the necessary data are being collected and that the evaluator (or evaluation team) knows exactly how he or she will attempt to reach conclusions about the program.

Different techniques are appropriate for different types of questions and evaluation purposes, and may depend on whether you have quantitative or qualitative data. Things to consider during the development of your plan include the following:

- How will responses be organized/ tabulated?
- Do you need separate tabulations for certain subgroups or program locations?
- What statistical techniques will be used?
- How will narrative data be analyzed?
- Who will do the analysis?
- How will the information be interpreted, and by whom? This is the process of attaching meaning to the analyzed data—numbers do not speak for themselves. Who should do it? Different perspectives can lead to different interpretations.

- What is the basis for interpreting the data? What criteria will guide the determination of what is meaningful?

It is important to keep in mind that data analysis and interpretation takes time, effort, and skill, usually more than you expect. Don't scrimp on this part of the evaluation—if necessary, cut back to a smaller study. The aim of analysis is to make sense out of the information you collect.

#### Initial Data Checking

- As you collect your data, you should regularly check the completed data collection forms (e.g., surveys) to make sure that the information being provided is complete and accurate. This process should include answering the following questions:

#### ***Getting Your Data Ready For Analysis***

- ***Are the responses clear and legible?***
- ***Are all questions answered?***
- ***Are the responses complete?***
- ***Are responses within the expected range?***
- ***Is all the necessary identifying information present?***

In almost every evaluation, data quality is a critical issue, so make sure that you do not shortchange this step. If you have incorporated open-ended questions, you may also want to develop codes so that responses can be more easily analyzed later and add these codes to your data file.

#### Choosing a Way to Store Your Data

In most cases you will want to enter your data into an electronic database to facilitate later analysis. The first thing to consider is what software to use, given the many options available. You should select a software package that can handle the types of analyses you want to do, and one that you or a staff member is capable of using; that is, one that isn't more complex than necessary.

Spreadsheet programs are considered by many to be among the most useful databases for quantitative and most kinds of qualitative data. The extent of statistical analysis you can accomplish with these programs is somewhat limited, but the program will store your data efficiently and provide you with counts, sums, differences, and other straightforward information. Other options include database programs, which have greater capabilities than spreadsheet programs, but databases can be somewhat more difficult to use.

For more complex statistical analysis, you will need a more powerful program that can allow you to conduct more sophisticated statistical procedures, such as regression. A variety of commercial packages are available for this purpose. These programs are less useful for qualitative data analysis, however. Fortunately, a variety of computer packages are also available for organizing and analyzing text and other types of qualitative information.

The next step is to decide which data structure is best suited to your data. In most cases, you will want to create a separate data record for each study participant, organized by unique identification (ID) number. If you have multiple data sources for each participant (e.g., a survey and data from existing records), you will want to be able to link the data using the participant's ID code.

Next, you will need to develop a "code book" that describes how each data record is organized; that is, the sequence of the variables. And for each variable you will want to specify: a unique variable name (a shorthand way to refer to each variable), description (e.g., race/ethnicity), format (number, text), source (data collection instrument), and any other relevant information. This is an important tool for the analysts, and is a way to document the data set for possible use at a future time.

### Data Entry

The most common way to enter information is through direct key entry (other options include data scanning and Web-based surveys). To ensure a high level of data accuracy, it is important to use double key entry—each data collection form is key entered twice and compared for any discrepancies arising through the data entry process.

### Missing Data and Data Transformations

The last step is to do some computer-assisted checking of your data. This will include checking the range of responses, checking the internal consistency of the responses, and examining missing data. How you decide to deal with any problems will depend on a variety of factors. Can you contact the respondent? If so, can you fix the problem? Are

there other data sources that you can use? There will likely be some problems that you will be unable to resolve. You will, therefore, have to decide to drop problem cases from your analysis (it may only be for certain analyses), or to “impute” data—that is, make a correction based on either some rule that is reasonable and sensible, or use “average” values from other similar respondents to replace missing values (there are many more complex statistical procedures for dealing with such problems).

Some data elements will also require transformation for use in analysis. For example, you may want to combine variables to create a new measure such as hours per year of training received.

## Types of Analysis

There are two major types of analytic procedures:

- **Descriptive statistics.** These can be as simple tables of the frequency of different responses to your questions, which you may also want to tabulate separately for different groups of interest. Other types of descriptive statistics include means (averages) and medians to show how typical something is, or measures of variability, such as ranges and standard deviations.
- **Inferential statistics** This type of statistical analysis, involves trying to draw conclusions about the extent to which any observed differences (e.g., between two or more groups of individuals or programs) are “statistically significant” or could have occurred by chance alone. Other, more complex, analyses (e.g., multivariate regression) seek to determine the statistical significance of a difference between groups *after controlling for any differences between the groups* (i.e., they are held constant).

Any analysis effort should begin by “getting to know the data” through the use of frequency distributions and graphical displays of the data. These early data explorations will be a critical step to understanding the information you have collected and what analysis it can or cannot support.

Next, you should move to a variety of descriptive statistics and simple inferential statistical procedures (e.g., *t*-tests of differences between means) before embarking on anything more sophisticated. Analysts often want to rush to the “fancy” analyses (modern computers and software make this far too easy to do) before they really have a grasp of their data. In many cases, the simple statistics will tell the main story and are generally easier to convey to your audience.

***In addition to using statistical tools, it is also important to interpret the findings, especially with regard to the results' practical significance.***

For example, one can find a statistically significant difference in, for example, test scores between two groups of children, but the magnitude of the estimated program impact may be too small to be meaningful from an educational perspective.

The hardest part, of course, is finding that story in the data. We suggest that you give yourself enough time to review the data thoughtfully, and that you discuss your findings with others—including those who collected the data—who can provide their insights and impressions of what the information means. You may need to review the information needs of your stakeholders and the goals of the program (including how these goals may have changed) to decide what is relevant to this evaluation.

## Getting the Story Down: Kramer School District Continued

The data are all collected, so now what? It is time to try to answer the research questions that Kramer School District posed at the beginning of its study.

Recall that the evaluation team had several questions that it wanted to answer:

- Does the summer institute increase teacher knowledge of the standards?
- Does the summer institute increase teacher knowledge of “best practices” in instruction?
- Does the summer institute improve the quality of teachers’ classroom instruction?
- Does the mentoring program increase teachers’ knowledge of student learning styles and how to deal with them?
- Does professional development lead to more positive teacher attitudes and greater motivation?
- Do the students of trained teachers exhibit higher engagement in the classroom?
- Are the students of trained teachers more likely to be engaged in cooperative learning?
- Are student outcomes “better” for teachers who participated in one of the components? in both components?

Although a discussion of the complete analysis plan for the study would be too lengthy for this handbook, some examples will help illustrate the key steps in the analysis process.

### Response Rates and Missing Data

Kramer first checked for nonresponse problems, including both data that were missing for complete individuals (teachers or students) and “item nonresponse,” which was due to respondents failing to answer one or more questions. In particular, the district looked for (a) high levels of nonresponse or missing data, and (b) systematic patterns of nonresponse or missing data.

With regard to the first issue, a district like Kramer should have at least a 70 percent response rate, which in this example means having at least one complete data point for 70 percent of the teachers and students *across the multiple data collection points*. If this were a single-shot survey, then Kramer would want at least a 70 percent response rate. (Longitudinal studies such as the one used in this example pose a greater challenge, because it is hard to maintain cooperation over time.) As for the second issue, Kramer wanted to check for systematic differences across its different study groups and for specific questions that may have been skipped (such as those appearing at the bottom of a page).

The Kramer evaluation had high response rates (over 80 percent) for teacher and administrator questionnaires. Upon receipt, all the surveys were carefully reviewed for missing data, and in most cases all questions were answered. In addition, in this example teachers were distinguished by the program component they were assigned to (including the comparison group), so the evaluation team made sure that there were no systematic response differences among the various study groups—which could have been mistaken for actual differences in program effects.

### Getting to Know the Data

District staff may be tempted to jump into analysis and get right down to trying to answer the research question. They shouldn't! Time spent up front getting to understand the data will be invaluable, saving many false steps and reducing the likelihood of errors in analysis and interpretation.

Good evaluators always begin by running basic descriptive statistics using their data, and then move on to various cross-tabulations to see how the different elements of the data

relate to one another. For example, the Kramer evaluation team produced basic statistics for each question in the their survey and classroom observations (e.g., means, medians, standard deviations, ranges) for each wave of data collection. They then produced various cross-tabulations: between survey items and comparing observation and survey data; comparisons across schools; comparisons by different teacher characteristics; and across the different study groups (e.g., mentoring only vs. mentoring plus the summer institute).

### Descriptive Analysis

In many cases, the research questions can be answered by using relatively simple descriptive analysis techniques. For example, the Kramer School District team used the classroom observations to create an overall rating of instructional quality for each teacher. These data were then tabulated for the different study groups both by year and as an overall pre-post change score (i.e., the change in instructional quality over the three time points). Statistical tests were then used to determine if any observed differences between groups were “statistically significant”—that is, was the difference likely to have come about by chance alone?

### Linking Outcomes to Program Activities

The final type of analysis done by the Kramer team was to try to determine if there were changes in student test scores that could be attributed to the teacher professional development. They knew this was a complex analysis issue, so they brought in Dr. Frank and his graduate students, who came up with a regression model that estimated the effect of receiving the professional development options on the average class-level test scores, controlling for the characteristics of the teacher and of the students in the class. Because the team did not use a randomized experiment, they needed to use this type of “statistical control” to try to eliminate other factors that may have “caused” changes in student test scores.

### Synthesizing Qualitative and Quantitative Data

Data relevant to each question were examined together to see how, for example, the effects on teachers could be explained through teacher questionnaires, school administrator interviews, and classroom observations. Wherever possible, the team compared data collected using different methods to see if one method of observation confirmed the results of another. For example, in examining effects on students, the team compared student achievement data with classroom observation data on student behavior. In addition, the team attempted to confirm (or triangulate) qualitative self-

reported data, such as that provided in interviews with school administrators, with data from the classroom observations. Such comparisons were simplified because the evaluation team had linked each item in their data collection instruments with a specified purpose related to the research questions (as discussed in chapter 7).

## Chapter 11. The Evaluation Report: Telling the Story

The students had a great time on their field trip and learned many new and exciting things. They are eager to share the experience with their parents and schoolmates. It is the same in an evaluation: Once you have collected the data and completed your analysis, you will want to disseminate your findings and, hopefully, use them to make decisions about your professional development program.

***How well you tell the story will, to a large extent, determine your ability to influence the decision-making process.***

### Interpretation of Results

***No matter how complex the evaluation and the analysis, the main goal is to tell a story! You will grab the attention of your audience if you can put together a clear, concise, and compelling story from your data.***

Although your evaluation report will provide the answers to your research questions, it should also convey the program's beginnings, its successes and challenges, and the specific nuances that made this program turn out the way it did—as well as the results of your study. Where possible (and if your design allows for it), the report should include examples, or vignettes, that support some of your conclusions.

Your use of various evaluation techniques should not be the defining feature of your report. Instead, try to take the perspective of an objective but genuinely interested observer, and attempt to tell the story of the program as richly as possible. By combining results seamlessly into an organic account of what you learned, your report will inspire confidence and actually *demonstrate* the soundness of your conclusions.

### Reporting

***The best advice for writing your report is to have an outline and start early.***

You should know where you are headed from the beginning; that is, you should already know your audience and their information needs, have research questions that will define what it is that you are trying to understand, and have an analysis plan that allows you to answer these questions. This knowledge should help you prepare a report outline well in

advance, and you should at least try to develop data displays (such as table shells and graphics) that are relevant for different research questions. Such advance planning will greatly facilitate the report-writing process.

Beyond advance planning, you should try to write up report sections early that can be prepared ahead of time. For example, the program description, evaluation objectives, and methodology sections can be written up at any time as background information for the report. Also, the notes you have collected on observations, interviews, and so on should be written up as soon as possible. Furthermore, some of the findings can be presented in a preliminary manner to help the evaluation team begin to think about conclusions and further areas for research.

### The Need for Multiple Products

You will definitely want to present your findings objectively and professionally. However, this does not mean that you should present the results uniformly to all stakeholders. Your stakeholders have varying interests in the findings, and there's no reason to provide them with the full set of findings (although it's always a good idea to give everyone the *option* to look at the full report). Rather, you should try to highlight those findings in which they expressed the most interest.

In addition, a thoughtful presentation of your results will help your audience get as much as possible out of the evaluation process. The more you target your written report or oral presentation to the needs of your audience, the more likely it is that they will pay attention to what the evaluation results suggest, which may encourage them to use the evaluation results more effectively, the subject of the next section of this chapter.

In addition, it will be important to report both positive and negative findings, as well as inconclusive findings. You must also indicate the limitations of the study design and any circumstances or events that hindered the implementation of the evaluation. In this way you will communicate how much confidence your audience can have in the results, and you will probably be able to shape their understanding of the constraints under which evaluations are conducted. By taking this careful approach, you may also help minimize their fear of evaluations in the future—and maximize the likelihood that they will support future evaluations to determine how best to improve the programs that can benefit from improvements.

Ideally, you should try to communicate that evaluation is a cumulative process, with your findings representing only one of many factors that contribute to decisions about the program.

#### Early Review of Your Results

***We suggest that you give program staff and others who participated in the evaluation a chance to review those sections to which they contributed. We also recommend that you brief key stakeholders on your results in advance of releasing the final report.***

Early review will allow you to correct omissions, misinterpretations, and other factual issues. Furthermore, after reviewing what you've written, these staff members may think of other insights and issues to discuss with you—ideas that did not occur to them during the course of the data collection.

You may also want to ask the program staff and other key stakeholders to review your preliminary results at some point before you complete the final report. Briefing high-level stakeholders about your report before making it public gives key officials a chance to think about the findings and their implications for the program and to prepare a response if they so desire. By circulating an interim report that describes your results, you may hear some new perspectives from your colleagues. These ideas may help you write your final report or cause you to reconsider some of your findings in light of the new information.

Note, however, that the purpose of such early reviews is *not* to mitigate so-called “negative” results. When disseminating preliminary findings to evaluation team members, you are trying to strengthen your analysis and be as accurate as possible. You might also discover that you need to go back and analyze your data in new ways or collect more data (you should do that if possible). When providing early results to stakeholders, you are giving them time to think about the implications of these findings and strategize about “next steps.” It is your responsibility to stand by your findings, however, and not succumb to any indirect pressures to “tone down” or “revise” actual results.

## Options for Presenting Your Results

***You have three basic options for reporting your results: a full report, an oral briefing, and an abbreviated report.***

You will want to prepare a **full report** after completing an evaluation. You should have written up various sections of the report in advance, particularly the background and methodology. However, much will need to be written at the end, after the data have been analyzed and after you have listened to feedback from some key stakeholders about the findings. The following are some basic elements of a final report.

### **Exhibit 8: Sections of the Final Report**

- ***Review the program***, providing background information about the program history, targeted audience(s), objectives, and activities. The earlier work of carefully defining the program will be helpful here. The logic model can be included as an efficient way of describing the program. This part of the report can be written even before the data collection has begun.
- ***Describe the reasons for conducting the evaluation***, including the key stakeholders' information needs and the final objectives for the evaluation (the research questions).
- ***Include a concise description of the evaluation methods*** (data collection and analysis) in the report and put more detailed information in a technical appendix.
- ***Describe findings*** (results), including positive, negative, and inconclusive findings, and offer interpretations of these findings. It may be useful to write a short summary of the conclusions organized by your evaluation questions. Then you can provide full documentation (e.g., data collection methods, response rates) for all of the findings in an appendix.
- ***Draw conclusions*** based on findings, and, if appropriate, supply recommendations for program improvement. In most cases, it is permissible to include conclusions that are based on a more subjective or intuitive understanding of the program gained from extended contact throughout the evaluation process. However, these more speculative conclusions—which can be based on anecdotal data—should be put at the end of the section, after you present conclusions based on hard data. Recommendations generally express views based on the evaluation team's overall impression of the program—and some may be creative and rely less on the data that were collected. Wherever possible, refer to data with phrases such as “as was suggested by several participants.”
- ***Provide an executive summary*** (one to four pages) to put at the beginning of the report. The executive summary is really a shortened version of the report, with a very short section on the program, evaluation objectives, and methods, and a succinct description of the key findings. The evaluator ought to keep in mind that many people, no matter how well-meaning, may read only the executive summary. Focus this section on the key findings and conclusions of the report. In addition, keep in mind that the conclusion and recommendations of the full report are the sections that may be read first, with the rest merely skimmed.
- Include a ***technical appendix*** with more precise methodological information, such as data collection instruments, and any additional data not presented in the report.

Options for shorter, more tailored reports include **oral briefings**, which are especially useful for individuals who are too busy to read the entire report. Also, an oral briefing can be given on the occasion of a visit from an important stakeholder, such as a state department of education official. Finally, those stakeholders with limited interest in the evaluation might appreciate a tailored presentation rather than the full report.

In devising the presentation, you should think carefully about your audience's information needs. What should and should not be included to best meet the needs of your particular audience? How should the presentation be organized? If you are talking to teachers, for example, you may wish to describe in detail the data you collected about teachers' reactions to the professional development they received, including information that reflects important perspectives from a teacher's point of view. Breaking this information down by grade level may also be useful. If you are speaking with the superintendent, you may be more interested in focusing on data regarding changes in instruction and the results of your student achievement analysis.

Similarly, you may decide that an **abbreviated report** is best suited to a particular audience. The school board may be less interested in details about implementation and more interested in final outcomes, for example. A well-crafted **executive summary** can also be disseminated as a short abbreviated report that will probably cover the information needs of most individuals.

#### Key Features of Well-Written Reports

High-quality evaluation reports share several characteristics. First, they are *succinct*. Your readers will be most interested in three things:

***What Readers Want To Know:***

- ***What was learned?***
- ***How reliable is this information?***
- ***What are the implications of this information for the future of this program and for the district overall?***

Inevitably, then, you will *not* include every detail or every finding in your report. But how do you decide what to include? The findings section could well be organized around the evaluation questions. This provides you with an organizational framework that will help limit the information in the report. We recommend looking at your evaluation questions and

including all information that answers these questions. You can then include a short section summarizing other findings that go beyond the evaluation questions but are of interest.

You also should ensure that the information you include is of interest to your key stakeholders. For example, details about evaluation methods should be presented only to demonstrate the degree of rigor in your work. They are not intrinsically very interesting (except maybe to other evaluators!) and should be kept as brief as possible. And if you cannot think of anyone who would want to know the information you're about to describe, think twice about including it.

The report, then, is not intended to be a comprehensive description of every impression and finding gained along the way. It is intended to respond to the evaluation objectives by answering the evaluation questions as completely as possible—without becoming obscure.

Good reports are also *lively*. It may be a challenge, but if you can write your report with a tone that is engaging and inquisitive, so that the reader is encouraged to wonder along with you about results of your team's investigations, then your report will be more of a pleasure to read. Try to picture your audience as you write. What would interest these individuals? What perspective would make sense to them? How should these details or findings be organized to get them to follow along?

But the reality is that some of what you need to convey is tedious. To relieve the reader, you should consider the following:

***To improve your report layout, try to use***

- ***Clear, specific quotes to enliven the text. Particularly memorable quotes can also be enlarged and set off on the page.***
- ***Visuals, such as graphs, tables, and diagrams, to display data trends.***
- ***Formatting to break up the text.***
- ***Graphics to make your descriptions more concrete.***

## Where Do You Go from Here?

***Once you have presented your report to a variety of stakeholders, most likely you will want to ensure that the recommendations that you made are considered seriously by district staff. One way to do this is to create an action plan.***

After the report is released, you can interview key stakeholders who are familiar with the findings. Ask them if they have reviewed the report, and offer to provide them with another copy if they have not. If they have read the report, ask them how they responded to its conclusions and recommendations. Ask them for suggestions about what could be done to address the remaining issues. Find out who is responsible for making changes in the program based on the evaluation. Although you may need to be patient, your stakeholders will probably appreciate your determination. Ideally, you will be able to bring key decisionmakers and stakeholders together to review the evaluation and develop an action plan outlining next steps for the program. Although things may not occur in exactly this manner, your goal is to get key decisionmakers to address the key findings and conclusions of the evaluation.

### Follow-Up Research

A single evaluation does not answer all the questions about any program. You will find that your audience has additional questions about the program, suggesting areas for further research. If the professional development program continues—even in a modified form—follow-up research can contribute to a more complete picture of how the schools are responding to efforts to develop instructional capacity. It may be that some questions could not be answered given the constraints on your evaluation, as was the case for Ringwood School District. Constraints could include the frame for your evaluation, funding, staff, or opportunities to collect data. To complete your story of the professional development program, you will undoubtedly need to conduct further evaluations.

### Use What You Learned

Your goal has been to understand the effectiveness of your professional development program. As with a field trip, you have taken a journey and now must ask, with your students, “What can we do with what we learned?”

If you answered formative questions, then your results provide evidence of how the program has been implemented. You can use this information to determine whether the program needs additional support (e.g., more teacher training, more instructional materials, more teacher incentives), and, if so, how to provide support that meets the needs of program participants. Your evaluation might also suggest that the program has been implemented as planned. Knowing this is also useful because it rules out any possibility that the results of an impact evaluation are related to insufficient implementation. In other words, you would know, with certainty, that a lack of impact of your program on student achievement is *not* because the program was not implemented fully. You could then consider other explanations, such as a mismatch between the goals of the professional development program and the material in the student assessment (suggesting a lack of alignment), or a deficiency in the training itself.

You can use the results of your impact evaluation to question or promote the continuation of a program. If the professional development is not yielding expected changes in student achievement—and you have determined that the time period for measuring such changes was sufficient—then the district will want to consider new options for staff development. If the program served a sample of teachers, it could be expanded to include other teachers. Including a different group of teachers might require modifying the program. Here, the results of any formative evaluation activities could guide you in making needed changes. And if the program has brought about expected changes, it too should be considered carefully to see whether it continues to serve a useful function.

You will want to include your evaluation results in requests for future or new funding. Even so-called “negative” results can be used to strengthen your argument for a new approach or an extended program. Your assurance to follow-up with more evaluation activities suggests a commitment to continuous improvement that will often be viewed as a plus by potential funders.

## Final Thoughts

We hope that you, like the teacher and students in our metaphorical field trip, have had a rewarding journey through the landscape of program evaluation. In most cases, we have just touched upon complex topics, but we have at least given you some tools that will help you learn more about your professional development program. The best way to learn is by doing—so just go do it!

And when you do, don't just let your evaluation report and presentations be all that's left in the district office months after the evaluation has been completed. With some extra effort, you can help your district improve its current programs as well as its capacity and willingness to continue funding useful evaluations.



## Appendix: Guiding Principles for Evaluators

---

The following are the principles established by the American Evaluation Association that guide professional evaluators. Keep them in mind and use them as guideposts in planning your evaluation:

### Systematic Inquiry

Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.

That is:

- Evaluators should adhere to the highest appropriate technical standards in conducting their work, whether that work is quantitative or qualitative in nature, so as to increase the accuracy and credibility of the evaluative information they produce.
- Evaluators should explore with the client the shortcomings and strengths both of the various evaluation questions it might be productive to ask, and the various approaches that might be used for answering those questions.
- When presenting their work, evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret, and critique their work. They should make clear the limitations of an evaluation and its results. Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses that *significantly* affect the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.

### Competence

Evaluators provide competent performance to stakeholders. This means:

- Evaluators should possess (or, here and elsewhere as appropriate, ensure that the evaluation team possesses) the education, abilities, skills, and experience appropriate to undertake the tasks proposed in the evaluation.
- Evaluators should practice within the limits of their professional training and competence, and should decline to conduct evaluations that fall substantially outside those limits. When declining the commission or request as not feasible or appropriate, evaluators should make clear any significant limitations on the evaluation that might result. Evaluators should make every effort to gain the competence directly or through the assistance of others who possess the required expertise.
- Evaluators should continually seek to maintain and improve their competencies, in order to provide the highest level of performance in their evaluations. This continuing

professional development might include formal coursework and workshops, self-study, evaluations of one's own practice, and working with other evaluators to learn from their skills and expertise.

## Integrity/Honesty

Evaluators ensure the honesty and integrity of the entire evaluation process. That is:

- Evaluators should negotiate honestly with clients and relevant stakeholders concerning the costs, tasks to be undertaken, limitations of methodology, scope of results likely to be obtained, and uses of data resulting from a specific evaluation. It is primarily the evaluator's responsibility to initiate discussion and clarification of these matters, not the client's.
- Evaluators should record all changes made in the originally negotiated project plans, and the reasons why the changes were made. If those changes would significantly affect the scope and likely results of the evaluation, the evaluator should inform the client and other important stakeholders in a timely fashion (barring good reason to the contrary, before proceeding with further work) of the changes and their likely impact.
- Evaluators should seek to determine, and where appropriate be explicit about, their own, their clients', and other stakeholders' interests concerning the conduct and outcomes of an evaluation (including financial, political and career interests).
- Evaluators should disclose any roles or relationships they have concerning whatever is being evaluated that might pose a significant conflict of interest with their role as an evaluator. Any such conflict should be mentioned in reports of the evaluation results.
- Evaluators should not misrepresent their procedures, data or findings. Within reasonable limits, they should attempt to prevent or correct any substantial misuses of their work by others.
- If evaluators determine that certain procedures or activities seem likely to produce misleading evaluative information or conclusions, they have the responsibility to communicate their concerns, and the reasons for them, to the client (the one who funds or requests the evaluation). If discussions with the client do not resolve these concerns, so that a misleading evaluation is then implemented, the evaluator may legitimately decline to conduct the evaluation if that is feasible and appropriate. If not, the evaluator should consult colleagues or relevant stakeholders about other proper ways to proceed (options might include, but are not limited to, discussions at a higher level, a dissenting cover letter or appendix, or refusal to sign the final document).
- Barring compelling reason to the contrary, evaluators should disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

## Respect for People

Evaluators respect the security, dignity and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact:

- Where applicable, evaluators must abide by current professional ethics and standards regarding risks, harms, and burdens that might be engendered to those participating in the evaluation; regarding informed consent for participation in evaluation; and regarding informing participants about the scope and limits of confidentiality. Examples of such standards include federal regulations about protection of human subjects, or the ethical principles of such associations as the American Anthropological Association, the American Educational Research Association, or the American Psychological Association. Although this principle is not intended to extend the applicability of such ethics and standards beyond their current scope, evaluators should abide by them where it is feasible and desirable to do so.
- Because justified negative or critical conclusions from an evaluation must be explicitly stated, evaluations sometimes produce results that harm client or stakeholder interests. Under this circumstance, evaluators should seek to maximize the benefits and reduce any unnecessary harm that might occur, provided this will not compromise the integrity of the evaluation findings. Evaluators should carefully judge when the benefits from doing the evaluation or in performing certain evaluation procedures should be foregone because of the risks or harms. Where possible, these issues should be anticipated during the negotiation of the evaluation.
- Knowing that evaluations often will negatively affect the interests of some stakeholders, evaluators should conduct the evaluation and communicate its results in a way that clearly respects the stakeholders' dignity and self-worth.
- Where feasible, evaluators should attempt to foster the social equity of the evaluation, so that those who give to the evaluation can receive some benefits in return. For example, evaluators should seek to ensure that those who bear the burdens of contributing data and incurring any risks are doing so willingly, and that they have full knowledge of, and maximum feasible opportunity to obtain any benefits that may be produced from the evaluation. When it would not endanger the integrity of the evaluation, respondents or program participants should be informed if and how they can receive services to which they are otherwise entitled without participating in the evaluation.
- Evaluators have the responsibility to identify and respect differences among participants, such as differences in their culture, religion, gender, disability, age, sexual orientation and ethnicity, and to be mindful of potential implications of these differences when planning, conducting, analyzing, and reporting their evaluations.

## Responsibilities for General and Public Welfare

Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare:

- When planning and reporting evaluations, evaluators should consider including important perspectives and interests of the full range of stakeholders in the object being evaluated. Evaluators should carefully consider the justification when omitting important value perspectives or the views of important groups.
- Evaluators should consider not only the immediate operations and outcomes of whatever is being evaluated, but also the broad assumptions, implications and potential side effects of it.
- Freedom of information is essential in a democracy. Hence, barring compelling reason to the contrary, evaluators should allow all relevant stakeholders to have access to evaluative information, and should actively disseminate that information to stakeholders if resources allow. If different evaluation results are communicated in forms that are tailored to the interests of different stakeholders, those communications should ensure that each stakeholder group is aware of the existence of the other communications. Communications that are tailored to a given stakeholder should always include all-important results that may bear on interests of that stakeholder. In all cases, evaluators should strive to present results as clearly and simply as accuracy allows so that clients and other stakeholders can easily understand the evaluation process and results.
- Evaluators should maintain a balance between client needs and other needs. Evaluators necessarily have a special relationship with the client who funds or requests the evaluation. By virtue of that relationship, evaluators must strive to meet legitimate client needs whenever it is feasible and appropriate to do so. However, that relationship can also place evaluators in difficult dilemmas when client interests conflict with other interests, or when client interests conflict with the obligation of evaluators for systematic inquiry, competence, integrity, and respect for people. In these cases, evaluators should explicitly identify and discuss the conflicts with the client and relevant stakeholders, resolve them when possible, determine whether continued work on the evaluation is advisable if the conflicts cannot be resolved, and make clear any significant limitations on the evaluation that might result if the conflict is not resolved.
- Evaluators have obligations that encompass the public interest and good. These obligations are especially important when evaluators are supported by publicly generated funds; but clear threats to the public good should never be ignored in any evaluation. Because the public interest and good are rarely the same as the interests of any particular group (including those of the client or funding agency), evaluators will usually have to go beyond an analysis of particular stakeholder interests when considering the welfare of society as a whole.

## References and Resources

---

### Professional Development

- Birman, B., et al. 1999. *Designing Effective Professional Development: Lessons From the Eisenhower Program*. Prepared for the U.S. Department of Education. Washington, D.C.: American Institutes for Research.
- Cohen, D., and H. Hill. 1998. *Instructional Policy and Classroom Performance: The Mathematics Reform in California*. (RR-39). Philadelphia, PA: Consortium for Policy Research in Education.
- Corcoran, T. 1995. *Helping Teachers Teach Well: Transforming Professional Development*. (CPRE Policy Brief). Philadelphia, PA: Consortium for Policy Research in Education.
- Corcoran, T., P. Shields, and A. Zucker. 1998. *The SSIs and Professional Development for Teachers: Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program*. Menlo Park, CA: SRI International.
- Darling-Hammond, L., and M. W. McLaughlin. 1995. "Policies That Support Professional Development in an Era of Reform." *Phi Delta Kappan* 76 (8): 597–604.
- Donnelly, M. B., T. Dove, and J. Tiffany. 2000. *Evaluation of Key Factors Impacting the Effective Use of Technology in Schools*. Prepared for the Planning and Evaluation Service, U.S. Department of Education. Arlington, VA: SRI International.
- Elmore, R., and R. Rothman, eds. 1999. *Testing, Teaching, and Learning: A Guide for States and School Districts*. A Report of the National Research Council. Washington, D.C.: National Academy Press.
- Horizon Research, Inc. 2000. "Information Regarding the Ongoing Evaluation of the NSF's Local Systemic Change Through Teacher Enhancement Program." See <http://www.horizon-research.com/LSC>.
- Kennedy, M. 1998. "Form and Substance in In-Service Teacher Education." Research Monograph #13. University of Wisconsin, Madison: National Institute for Science Education.
- Loucks-Horsley, S. 1998. "The Role of Teaching and Learning in Systemic Reform: A Focus on Professional Development." *Science Educator* 7 (1):1–6.
- Loucks-Horsley, S., P. W. Hewson, N. Love, and K. E. Stiles. 1998. *Designing Professional Development for Teachers of Science and Mathematics*. A project of the National Institute for Science Education. Thousand Oaks, CA: Corwin Press.

- Milbrey, M., and J. Talbert. 1993. "Introduction: New Visions of Teaching." In *Teaching for Understanding: Challenges for Policy and Practice*, edited by D.K. Cohen, M. W. McLaughlin, and J. Talbert. 1993. San Francisco: Jossey-Bass.
- Mitchell, A., and J. Raphael. 1999. *Goals 2000: Case Studies of Promising Districts*. A report prepared by The Urban Institute. Washington, D.C.: U.S. Department of Education.
- National Board for Professional Teaching Standards. 1989. See <http://www.nbpts.org/nbpts/standards>.
- National Center for Education Statistics. 1999. *Teacher Quality: A Report on the Preparation and Qualifications of Public School Teachers*. U.S. Department of Education, Office of Educational Research and Improvement. Washington, D.C.: U.S. Department of Education.
- National Council of Teachers of English. 1996. *NCTE/IRA Standards for the English Language Arts*. Urbana, IL: National Council of Teachers of English.
- National Council of Teachers of Mathematics. 1989. *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- National Council of Teachers of Mathematics. 1991. *Professional Standards for Teaching Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- National Partnership for Excellence and Accountability in Teaching. 1999. *Characteristics of Effective Professional Development*. See <http://www.npeat.org>.
- Sparks, D., and S. Hirsh. 2000. *A National Plan for Improving Professional Development*. Oxon, OH: National Staff Development Council.
- Shields, P., J. Marsh, and N. Adelman. 1998. *Evaluation of NSF's Statewide Systemic Initiatives (SSI) Program: The SSI's Impacts on Classroom Practice*. Menlo Park, CA: SRI, International.
- U.S. Department of Education. (n.d.). *Principles of High-Quality Professional Development*. Washington, D.C.: Government Printing Office.
- Wilson, S. M., and J. Berne. 1999. "Teacher Learning and the Acquisition of Professional Knowledge: An Examination of Research on Contemporary Professional Development." *Review of Research in Education* 24:173–209.

## Evaluating Professional Development

- Finnan, C., and S. C. Davis. 1995. "Linking Project Evaluation and Goals-Based Teacher Evaluation: Evaluating the Accelerated Schools Project in South Carolina." Paper presented at the Annual Meeting of the National Evaluation Institute, 18–22 April, in San Francisco, CA. ERIC document number 384 637.
- Guskey, T. R. 1999. "New Perspectives on Evaluating Professional Development." Paper presented at the Annual Meeting of the American Educational Research Association, 19–23 April, in Montreal, Canada. ERIC document number 430 024.
- Guskey, T. R. 2000. *Evaluating Professional Development*. Thousand Oaks, CA: Corwin Press.

- Mack, P. J. 1998. "A Spiral Plan for Delivery and Evaluation of Continuous Professional Development: A Guide Adapted from the Work of M. Fullan, D. Kirkpatrick, A. Costa, and B. Kallick." ERIC document number 426 981.
- Reed, C. J., F. K. Kochan, M. E. Ross, and R. C. Kukel. 1999. "Frameworks for Summative and Formative Evaluation of Diverse PDS Sites." Paper presented at the Annual Meeting of the American Educational Research Association, 19–23 April, in Montreal, Canada. ERIC document number 429 941.
- Roberts, L. 1997. "Evaluating Teacher Professional Development: Local Assessment Moderation and the Challenge of Multisite Evaluation." Paper presented at the Annual Meeting of the National Evaluation Institute, 9 July, in Indianapolis, IN. ERIC document number 413 360.
- Scannel, D. P. 1996. "Evaluating Professional Development Schools: The Challenge of an Imperative." *Contemporary Education* 67 (4): 241–243.
- Stein, M., J. Norman, and J. Clay-Chambers. 1997. "Assessing the Impact of an Urban Systemic Professional Development Program on Classroom Practice." Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Oak Brook, IL. ERIC document number 406 146.

## General Evaluation

- Administration for Children and Families. 1999. *The Program Manager's Guide to Evaluation*. Washington, D.C.: U.S. Department of Health and Human Services.
- American Evaluation Association. Guiding Principles For Evaluators. See <http://www.eval.org>.
- Atkinson, A.J., et al. 1999. *Program Planning and Evaluation Handbook: A Guide for Safe and Drug-Free Schools and Communities Act Programs*. Harrisonburg, VA: The Virginia Effective Practices Project, James Madison University.
- Berk, R.A., and P.H. Rossi. 1990. *Thinking About Program Evaluation*. Newbury Park, CA: Sage Publications.
- Bond, S.L., S.E. Boyd, K.A. Rapp, J.B. Raphael, and B.A. Sizemore. 1997. *Taking Stock: A Practical Guide to Evaluating Your Programs*. Chapel Hill, NC: Horizon Research, Inc.
- Campbell, D.T., and J.C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton-Mifflin.
- Card, J.J., C. Brindis, J.L. Peterson, and S. Niegro. 1999. *Guidebook: Evaluating Teen Pregnancy Prevention Programs*. Atlanta, GA: Centers for Disease Control and Prevention.
- Centers for Disease Control and Prevention. 1999. *Framework for Program Evaluation in Public Health*. Atlanta, GA: Centers for Disease Control and Prevention.
- Chen, H. 1990. *Theory-Driven Evaluations*. Newbury Park, CA: Sage Publications.

- Cicchinelli, L.F., and Z. Barley. 1999. *Evaluating for Success*. Aurora, CO: Mid-Content Research for Education and Learning.
- Denzin, N. K., and Y. S. Lincoln, eds. 1994. *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications.
- Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis For Field Settings*. Chicago, Illinois: Rand McNally.
- Cronbach, L. J., et al. 1980. *Toward Reform of Program Evaluation*. San Francisco, CA: Jossey-Bass.
- Frechtling, J., ed. 1995. *Footprints: Strategies for Non-Traditional Program Evaluation*. Arlington, VA: National Science Foundation.
- Frechtling, J., and L. Sharp, eds. 1997. *User-Friendly Handbook for Mixed Method Evaluations*. Arlington, VA: National Science Foundation.
- Freeman, H.E., P. H. Rossi, and G. D. Sandefur. *Workbook For Evaluation*. Newbury Park, CA: Sage Publications.
- Gray, S. T. 1993. *Leadership: A Vision of Evaluation*. Washington, D.C.: Independent Sector.
- Guskey, T. 1999. "New Perspectives on Evaluating Professional Development." Paper presented at the annual meeting of the American Educational Research Association, April, in Montreal, Canada.
- Harvey, T., ed. 1998. *Evaluation Cookbook*. Edinburgh, Scotland: Heriot-watt, Ltd.
- Hatry, H. P., and M. Kopczynski. 1997. *Guide to Program Outcome Measurement*. Prepared for the Planning and Evaluation Service, U.S. Department of Education. Washington, D.C.: The Urban Institute.
- Hedrick, T. E., L. Bickman, and D. J. Rog. 1993. *Applied Research Design*. Thousand Oaks, CA: Sage Publications.
- Herman, J., and L. Winters. 1992. *Tracking Your School's Success: A Guide to Sensible Evaluation*. Newbury Park, CA: Corwin Press.
- Herman, J., L. Morris, and C. Fitz-Gibbon. 1987. *Evaluator's Handbook*. Newbury Park, CA: Sage Publications.
- Learning Technology Dissemination Initiative. 1998. *Evaluation Cookbook*. Edinburgh, United Kingdom: Heriot-Watt University. See <http://www.icbl.hw.ac.uk/ltidi/cookbook/>.
- Light, R. J., J. D. Singer, and J.B. Willett. 1990. *By Design: Planning Research on Higher Education*. Cambridge, MA: Harvard University.
- Linney, J. A., and A. Wandersman. 1991. *Prevention Plus III: Assessing Alcohol and Other Drug Prevention Programs at the School and Community Level: A Four-Step Guide to Useful Program Assessment*. Washington, D.C.: Office of Substance Abuse Prevention, U.S. Department of Health and Human Services.

- Love, A. J. 1991. *Internal Evaluation: Building Organizations from Within*. Applied Social Research Methods Series, vol. 24. Newbury Park, CA: Sage Publications.
- Maxwell, J. A. 1996. *Qualitative Research Design: An Interactive Approach*. Thousand Oaks, CA: Sage Publications.
- McNabb, M., M. Hawkes, and R. Ullik. 1999. "Critical Issues in Evaluating the Effectiveness of Technology." Summary of the Secretary's Conference on Educational Technology: Evaluating the Effectiveness of Technology, July 12–13, 1999. Washington, DC. See <http://www.ed.gov/Technology/TechConf/1999/>.
- Miller, D. C. 1991. *Handbook of Research Design and Measurement*. Thousand Oaks, CA: Sage Publications.
- Mishan, E. J. 1988. *Cost-Benefit Analysis*. 4th ed. London: Unwin Hyman.
- Mohr, L. B. 1995. *Impact Analysis for Program Evaluation*. 2nd ed. Thousand Oaks, CA: Sage Publications.
- Muraskin, L. D. 1993. *Understanding Evaluation: The Way to Better Prevention Programs*. Washington, D.C.: U.S. Department of Education.
- Nagy, J., and S. B. Fawcett, eds. 1999. *Community Tool Box: Evaluating Comprehensive Community Initiatives*. Lawrence, KS: University of Kansas.
- . 1999. *Community Tool Box: A Framework for Program Evaluation*. Lawrence, KS: University of Kansas. Adapted from "Recommended Framework for Program Evaluation in Public Health Practice," by B. Milstein, S. Wetterhall, and the CDC Evaluation Working Group.
- Ontario Ministry of Health. 1999. *Program Evaluation Tool Kit*. Ottawa, Canada: Ontario Ministry of Health. See <http://www.uottawa.ca/academic/med/sld002.htm>.
- Patton, M. Q. 1997. *Utilization-Focused Evaluation*. 3rd ed. Thousand Oaks, CA: Sage Publications.
- . 1990. *Qualitative Evaluation and Research Methods*. 2nd ed. Beverly Hills, CA: Sage Publications.
- . 1987. *How to Use Qualitative Methods in Evaluation*. Newbury Park, CA: Sage Publications.
- . 1982. *Practical Evaluation*. Beverly Hills, CA: Sage Publications.
- Quinones, S., and R. Kirshtein. 1998. *An Educator's Guide to Evaluating the Use of Technology in Schools and Classrooms*. Washington, D.C.: U.S. Department of Education.
- Rossi, P.H., and H.E. Freeman. 1999. *Evaluation: A Systemic Approach*. 6<sup>th</sup> ed. Thousand Oaks, CA: Sage Publications.
- Sanders, J.R. 1992. *Evaluating School Programs: An Educator's Guide*. Newbury Park, CA: Corwin Press.
- Scriven, M. 1991. *Evaluation Thesaurus*. 4th ed. Newbury Park, CA: Sage Publications.

- Scriven, M. 1967. "The Methodology of Evaluation." In *Perspectives of Curriculum Evaluation*, edited by R. E. Stake. Chicago, IL: Rand-McNally.
- Shadish, W. R., T. D. Cook, and L. Leviton. 1991. *Foundations of Program Evaluation*. Newbury Park, CA: Sage Publications.
- Stake, R. E., ed. 1967. *Perspectives of Curriculum Evaluation*. Chicago, IL: Rand-McNally.
- Stecher, B. M., and W. A. Davis. 1987. *How to Focus an Evaluation*. Newbury Park, CA: Sage Publications.
- Stevens, F., et al. 1993. *User-Friendly Handbook for Project Evaluation in Science, Mathematics, Engineering, and Technology Education*. Arlington, VA: National Science Foundation.
- Taylor-Powell, E., B. Rossing, and J. Geran. 1998. *Evaluating Collaboratives: Reaching the Potential*. Madison, WI: Program Development and Evaluation, University of Wisconsin-Extension, Cooperative Extension.
- Taylor-Powell, E., S. Steele, and M. Douglah. 1996. *Planning a Program Evaluation*. Madison, WI: University of Wisconsin-Extension, Cooperative Extension.
- Trochim, W., and D. Land. 1982. "Designing Designs For Research." *The Researcher* 1 (1): 1–6.
- United Way of America. 1996. *Measuring Program Outcomes: A Practical Approach*. Alexandria, VA: United Way of America.
- Veney, J. E., and A. D. Kaluzny. 1998. *Evaluation and Decision-Making for Health Services*. 3rd ed. Ann Arbor, MI: Health Administration Press.
- W.K. Kellogg Foundation. 1998. *W.K. Kellogg Foundation Evaluation Handbook*. Battle Creek, MI: W.K. Kellogg Foundation.
- Worthen, B. R., and J. R. Sanders. 1987. *Educational Evaluation: Alternative Approaches and Practical Guidelines*. New York, NY: Longman.

## Collecting Data

- Denzin, N. K., and Y. S. Lincoln, eds. 1994. *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications.
- Fink, A. 1995. *The Survey Kit*. Thousand Oaks, CA: Sage Publications.
- Krueger, R. A. 1994. *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks, CA: Sage Publications.
- Marshall, C., and G. B. Rossman. 1995. *Designing Qualitative Research*. 2<sup>nd</sup> ed. Newbury Park, CA: Sage Publications.
- Patton, M. Q. 1990. *Qualitative Evaluation and Research Methods*. 2<sup>nd</sup> ed. Newbury Park, CA: Sage Publications.
- Stake, R. A. 1995. *The Art of Case Study Research*. Thousand Oaks, CA: Sage Publications.

Yin, R. K. 1994. *Case Study Research: Design and Methods*. Newbury Park, CA: Sage Publications.

## Data Analysis

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Bernhardt, V. 1998. *Data Analysis for Comprehensive Schoolwide Improvement*. Larchmont, NY: Eye on Education.

Goldberger, A.S. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Hsiao, C. 1986. *Analysis of Panel Data*. Cambridge, MA: Econometric Society Monograph No. 11, Cambridge University Press.

Jaeger, R. M. 1990. *Statistics: A Spectator Sport*. Newbury Park, CA: Sage Publications.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, MA: Econometric Society Monographs, Cambridge University Press.

Miles, M. B., and A. M. Huberman. 1994. *Qualitative Data Analysis—An Expanded Sourcebook*. 2<sup>nd</sup> ed. Thousand Oaks, CA: Sage Publications.

Mohr, L. B. 1995. *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: Sage Publications.

## Reporting Results

Morris, L. L., and C. T. Fitz-Gibbon. 1987. *How to Present an Evaluation Report*. Newbury Park, CA: Sage Publications.

Patton, M. Q. 1986. *Utilization Focused Evaluation*. 2<sup>nd</sup> ed. Newbury Park, CA: Sage Publications.

Weiss, C. H. 1988. "Evaluation for Decisions: Is Anybody There? Does Anybody Care?" *Evaluation Practice* 9 (5): 5–21.

## Useful Web Sites

Trochim, W. M. 1999. *The Research Methods Knowledge Base*, 2<sup>nd</sup> Edition. <http://trochim.human.cornell.edu/kb/index.htm>.

Mid-continent Research for Education and Learning (McRel). <http://www.mcrel.org>.

U.S. Department of Education. <http://www.ed.gov>.

Horizon Research, Inc.'s Local Systemic Change through Teacher Enhancement program evaluation: <http://www.horizon-research.com/LSC/>

## Glossary of Common Evaluation Terms

---

**Accountability**—The responsibility for the justification of expenditures, decisions, or the results of one's own efforts.

**Achievement**—A manifested performance determined by some type of assessment or testing.

**Anonymity** (provision for)—Evaluator action to ensure that the identity of subjects cannot be ascertained during the course of a study, in study reports, or any other way.

**Attrition**—Loss of subjects from the defined sample during the course of a longitudinal study.

**Audience(s)**—Consumers of the evaluation; that is, those who will, or should, read or hear of the evaluation, either during or at the end of the evaluation process. Includes those persons who will be guided by the evaluation in making decisions and all others who have a stake in the evaluation (see stakeholders).

**Background**—The contextual information that describes the reasons for the project, its goals, objectives, and stakeholders' information needs.

**Baseline data**—Initial information on program participants or other program aspects collected prior to receipt of services or program intervention. Baseline data are often gathered through intake interviews and observations and are used later for comparing measures that determine changes in your participants, program, or environment.

**Bias** (sample)—Error due to nonresponse or incomplete response from selected sample subjects.

**Bias** (statistical)—Inaccurate representation that produces systematic error in a research finding. Bias may result in over- or underestimation of certain characteristics of the population. It may result from incomplete information or invalid collection methods and may be intentional or unintentional.

**Case study**—An intensive, detailed description and analysis of a single project, program, or instructional material in the context of its environment.

**Client**—The person or group or agency that commissioned the evaluation.

**Coding**—The process of translating a given set of data or items into machine-readable categories.

**Cohort**—A term used to designate one group among many in a study. For example, "the first cohort" may be the first group to have participated in a training program.

**Comparison group**—Individuals whose characteristics (such as race/ethnicity, gender, and age) are similar to those of program participants. These individuals may not receive any services, or they may receive a different set of services, activities, or products. In no instance do they receive the same service(s) as those being evaluated. As part of the evaluation process, the treatment (or experimental) group and the comparison group are assessed to determine which type of services, activities, or products provided by a program produced the expected changes.

**Conclusions** (of an evaluation)—Final judgments and recommendations.

**Content analysis**—A process using a parsimonious classification system to determine the characteristics of narrative text.

**Control group**—A type of comparison group in which participants are randomly assigned to either the treatment (or experimental) group or the control group. The function of the control group—as with any comparison group—is to determine the extent to which the same effect occurs without the treatment. But control groups have been manipulated experimentally, usually by random assignment, to ensure that they are equivalent to the treatment group.

**Correlation**—A statistical measure of the degree of relationship between two or more variables.

**Cost analysis**—The process of calculating the cost of something that is being evaluated. Cost analyses usually try to determine (1) the relevant cost elements (e.g., staff labor, travel), (2) who bears the costs (i.e., who pays), and (3) the time period over which costs are incurred.

**Cost-effectiveness analysis**—A type of analysis that involves comparing the relative costs of operating a program with the extent to which the program met its goals and objectives. For example, a program to reduce cigarette smoking would estimate the dollars that had to be expended in order to convert each smoker into a nonsmoker.

**Cost-benefit analysis**—A type of analysis that involves comparing the relative costs of operating a program (program expenses, staff salaries, etc.) to the monetary benefits (gains to individuals or society) it generates. For example, a program to reduce cigarette smoking would focus on the difference between the dollars expended for converting smokers into nonsmokers with the dollar savings from reduced medical care for smoking related disease, days lost from work, and the like.

**Criterion-referenced test**—Tests whose scores are interpreted by referral to well-defined domains of content or behaviors, rather than by referral to the performance of some comparable group of people.

**Cross-case analysis**—A type of analysis comparing and contrasting results across separate case studies.

**Cross-sectional study**—A cross-section is a random sample of a population, and a cross-sectional study examines this sample at one point in time. Successive cross-sectional studies can be used as a substitute for a longitudinal study.

**Cultural relevance (or competency or sensitivity)**—Demonstration that evaluation methods, procedures, and or instruments are appropriate for the culture(s) to which they are applied.

**Data**—Specific information or facts that are collected. A data item is usually a discrete or single measure. Examples of data items might include age, date of entry into program, or reading level. Sources of data may include case records, attendance records, referrals, assessments, interviews, and the like.

**Data analysis**—The process of systematically applying statistical and logical techniques to describe, summarize, and compare data collected.

**Data collection instruments**—Forms used to collect information for an evaluation. Forms may include interview instruments, intake forms, case logs, and attendance records. They may be developed specifically for an evaluation or modified from existing instruments. A professional evaluator can help select those that are most appropriate for a given program.

**Data collection plan**—A written document describing the specific procedures to be used to gather the evaluation information or data. The plan describes who collects the information, when and where it is collected, and how it is to be obtained.

**Data reduction**—Process of selecting, focusing, simplifying, abstracting, and transforming data collected in written field notes or transcriptions.

**Database**—An accumulation of information that has been systematically organized for easy access and analysis. Databases typically are computerized.

**Dependent variable**—The variable that represents an outcome of interest (e.g., student achievement). The opposite of dependent variables are independent variables (e.g., student's age, family income), some of which can be manipulated (e.g., student has a teacher who has received professional development training).

**Descriptive statistics**—Statistics that involve summarizing, tabulating, organizing, and graphing data for the purpose of describing objects or individuals that have been measured or observed.

**Design**—The overall plan and specification of the approach expected in a particular evaluation. The design describes how program components will be measured and how the resulting measurements will be used. A pre- and post-intervention design with or without a comparison or control group is the design needed to evaluate participant outcome objectives.

**Dissemination**—The process of communicating information to specific audiences for the purpose of extending knowledge, sometimes with a view to modifying policies and practices.

**Evaluation**—A systematic method for collecting, analyzing, and using information to answer basic questions about your program. It helps to identify effective and ineffective services, practices, and approaches.

**Evaluation plan**—A written document describing the overall approach or design that will guide an evaluation. It includes what researchers plan to do, how they plan to do it, who will do it, when it will be done, and why the evaluation is being conducted. The evaluation plan serves as a guide for the evaluation.

**Evaluation team**—The individuals, such as the outside evaluator, evaluation consultant, program manager, and program staff who participate in planning and conducting the evaluation. Team members assist in developing the evaluation design, developing data collection instruments, collecting data, analyzing data, and writing the report.

**Evaluator**—An individual trained and experienced in designing and conducting an evaluation that uses tested and accepted research methodologies.

**Executive summary**—A nontechnical summary statement designed to provide a quick overview of the full-length report on which it is based.

**Experimental design**—The plan of an experiment, including selection of subjects who receive treatment and control group (if applicable), procedures, and statistical analyses to be performed.

**Experimental group**—See the definition for treatment group.

**External evaluation**—Evaluation conducted by an evaluator from outside the organization within which the object of the study is housed.

**Extrapolate**—To infer an unknown from something that is known. (Statistical definition—to estimate the value of a variable outside its observed range.)

**False negative**—Also called a Type II error, a statistical occurrence when an event that is not predicted actually occurs.

**False positive**—Also called a Type I error, a statistical occurrence when an event that is predicted occurs.

**Feasibility**—The extent to which an evaluation is appropriate for implementation in practical settings.

**Field test**—The study of a program, project, or instructional material in settings like those where it is to be used. Field tests may range from preliminary primitive investigations to full-scale summative studies.

**Focus group**—A group of 7 to 10 people convened for the purpose of obtaining perceptions or opinions, suggesting ideas, or recommending actions. Use of a focus group is a method of collecting data for evaluation purposes.

**Formative (or process or implementation) evaluation**—An evaluation that examines the extent to which a program is operating as intended by assessing ongoing program operations and whether the targeted population is being served. A process evaluation involves collecting data that describe program operations in detail, including the types and levels of services provided, the location of service delivery, staffing, sociodemographic characteristics of participants, the community in which services are provided, and the linkages with collaborating agencies. A process evaluation helps program staff identify needed interventions and/or change program components to improve service delivery. It is also called formative or implementation evaluation.

**Gain scores**—The difference between a student's performance on a test and his or her performance on a previous administration of the same test, such as the difference between fall and spring testing.

**Generalizability**—The extent to which information about a program, project, or instrumental material collected in one setting (e.g., a particular school receiving teacher training) can be used to reach a valid judgment about how it will perform in other settings (e.g., other schools in the district).

**Hawthorne effect**—The tendency of a person or group being investigated to perform better (or worse) than they would in the absence of the investigation, thus making it difficult to identify treatment effects.

**Hypothesis testing**—The standard model of the classical approach to scientific research, in which a hypothesis is formulated before the experiment to test its truth. The results are stated in probability terms that the results were due solely to chance. The significance level of 1 chance in 20 (.05) or 1 chance in 100 (.01) is a high degree of improbability.

**Immediate outcomes**—The changes in program participants, knowledge, attitudes, and behavior that occur early in the course of the program. They may occur at certain program points or at program completion. For example, changing teacher attitudes or knowledge as a result of professional development are an immediate outcome.

**Impact evaluation**—See the definition for summative evaluation.

**Implementation evaluation**—See the definition for formative evaluation.

**In-depth interview**—A guided conversation between a skilled interviewer and an interviewee that seeks to maximize opportunities for the expression of a respondent's feelings and ideas through the use of open-ended questions and a loosely structured interview guide.

**Indicator**—A factor, variable, or observation that is empirically connected with the criterion variable, a correlate. For example, judgment by students that a course has been valuable to them for pre-professional training is an indicator of the program's value.

**Inferential statistics**—These statistics are inferred from characteristics of samples to characteristics of the population from which the sample comes.

**Informed consent**—Agreement by the participants in an evaluation of the use of their names and/or confidential information supplied by them in specified ways, for stated purposes, and in light of possible consequences, made prior to the collection and/or release of this information in evaluation reports.

**Instrument**—A tool used to collect and organize information. Includes written instruments or measures, such as questionnaires, scales, and tests.

**Interaction**—A statistical concept in which the effect of one variable (e.g., teacher training) is hypothesized to vary with another variable (e.g., teacher experience). For example, the effect of teacher training is expected to depend upon the prior experience of the individual teachers (i.e., more experienced teachers get more out of the training).

**Intermediate outcomes**—Results or outcomes of a program or treatment that may require some time before they are realized. For example, improved classroom instruction would be an intermediate outcome of a program designed to increase student learning.

**Internal evaluator(s)**—The staff who conduct the evaluation and are part of the organization that is implementing the program or treatment.

**Intervention**—The specific services, activities, or products developed and implemented to change or improve program participants' knowledge, attitudes, behaviors, or awareness.

**Key informant**—Person with background, knowledge, or special skills relevant to topics examined by the evaluation.

**Level of significance**—The probability that the observed difference occurred by chance.

**Logic model**—See the definition for program model.

**Longitudinal study**—An investigation or study in which a particular individual or group of individuals is followed over a substantial period of time to discover changes that may be attributable to the influence of the treatment, maturation, or the environment.

**Management information system (MIS)**—An information collection and analysis system, usually computerized, that facilitates access to program and participant information. It is usually designed and used for administrative purposes. The types of information typically included in an MIS are service delivery measures, such as session, contacts, or referrals; staff caseloads; client sociodemographic information; client status; and treatment outcomes. Many MISs can be adapted to meet evaluation requirements.

**Matching**—An experimental procedure in which the subjects are divided, by means other than a lottery, so that the groups can be considered to be of equal merit or ability. (Matched groups are often created by ensuring that they are the same, or nearly so, on such variables as sex, age, grade point averages, and past test scores.)

**Mean**—Also called "average" or arithmetic average. For a collection of raw test scores, the mean score is obtained by adding all scores and dividing by the number of people taking the test.

**Measurement**—Determination of the magnitude of a quantity (e.g., a standardized test score is a measurement of reading achievement).

**Median**—The point in a distribution that divides the group into two, as nearly as possible. For example, in a score distribution, half the scores fall above the median and half fall below.

**Meta-analysis**—The name for a particular approach to synthesizing multiple quantitative studies on a common topic. It usually involves the calibration of a specific parameter for each study, called an “effect size.”

**Methodology**—The way in which information is discovered; a methodology describes how something will be (or was) done. The methodology includes the methods, procedures, and techniques used to collect and analyze information.

**Mixed-method evaluation**—An evaluation for which the design includes the use of both quantitative and qualitative methods for data collection and data analysis.

**Mode**—The value that occurs more often than any other. If all scores (in a score distribution) occur with the same frequency, there is no mode. If the two highest score values occur with the same frequency, there are two modes.

**Moderator**—Focus group leader; often called a focus group facilitator.

**Monitoring**—The process of reviewing a program or activity to determine whether set standards or requirements are being met. Unlike evaluation, monitoring compares a program to an ideal or exact state.

**“No significant difference”**—A decision that an observed difference between two statistics occurred by chance.

**Nominal data**—Data that consist of categories only, without order to these categories (i.e., region of the country, courses offered by an instructional program.)

**Norm**—A single value, or a distribution of values, constituting the typical performance of a given group.

**Norm-referenced tests**—Tests that measure the *relative* performance of the individual or group by comparison with the performance of other individuals or groups taking the same test.

**Objective**—A specific statement that explains how a program goal will be accomplished. For example, an objective of the goal to improve instructional practice could be to provide training to teachers on a weekly basis for six months. An objective is stated so that changes—in this case, an increase in specific skills—can be measured and analyzed. Objectives are written using measurable terms and are time-limited.

**Ordered data**—Non-numeric data in ordered categories (e.g., students’ performance categorized as excellent, good, adequate, and poor).

**Outcome**—A result of the program, services, or products provided, outcomes refer to changes in knowledge, attitude, or behavior in participants. Such changes are referred to as participant outcomes.

**Outcome evaluation**—See the definition for summative evaluation.

**Outcome objectives**—The changes in knowledge, attitudes, awareness, or behavior that you expect to occur as a result of implementing your program component, service, or activity. Also known as participant outcome objectives.

**Participant**—An individual, family, agency, neighborhood, community, or state receiving or participating in services provided by your program. Also known as a target population group.

**Peer review**—Evaluation done by a panel of judges with specific technical qualifications.

**Performance evaluation**—A method of assessing what skills students or other project participants have acquired by examining how they accomplish complex tasks or the products they have created (e.g., poetry, artwork).

**Pilot test**—A brief and simplified preliminary study designed to try out methods to learn whether a proposed project or program seems likely to yield valuable results (also called a pre-test).

**Population**—All persons in a particular group.

**Post-test**—A test to determine performance after the administration of a program, project, or instructional material.

**Pre-test**—A test to determine performance prior to the administration of a program, project, or instructional material. Pre-tests serve two purposes: diagnostic and baseline. This term can also refer to the use of an instrument (questionnaire, test, observation schedule) with a small group to detect need for revisions prior to use in a full-scale study.

**Process evaluation**—See the definition for formative evaluation.

**Program evaluation**—The systematic collection, analysis, and reporting of information about a program, to assist in decisionmaking.

**Program model (or logic model)**—A diagram showing the logic or rationale underlying a particular program. In other words, it is a picture of a program that shows what it is supposed to accomplish. A logic model describes the links between program objectives, program activities, and expected program outcomes.

**Purposive sampling**—Creating samples by selecting information-rich cases from which one can learn a great deal about issues of central importance to the purpose of the evaluation.

**Qualitative evaluation**—The approach to evaluation that is primarily descriptive and interpretative.

**Quantitative evaluation**—The approach to evaluation involving the use of numerical measurement and data analysis based on statistical methods.

**Quasi-experimental**—An evaluation design that seeks to approximate a true randomized experiment by identifying a group that closely matches the treatment or experimental group.

**Random assignment**—The assignment of individuals in the pool of all potential participants to either the experimental (treatment) or control group in such a manner that their assignment to a group is determined entirely by chance.

**Random sampling**—The process of drawing a number of items of any sort from a larger group or population so that every individual item has a specified probability of being chosen.

**Reliability**—Extent to which a measurement (e.g., an instrument or a data collection procedure) produces consistent results over repeated observations or administrations of the instrument under

the same conditions each time. It is also important that reliability be maintained across data collectors; this is called inter-rater reliability.

**Replication**—The process of repeating an intervention or evaluation with all essentials unchanged. Replications are often difficult to evaluate because of changes in design or execution.

**Response bias**—Error due to incorrect answers.

**Sample**—A subset of participants selected from the total study population. Samples can be random (selected by chance, such as every sixth individual on a waiting list) or nonrandom (selected purposefully, such as all third-grade students).

**Sampling error**—Error due to using a sample instead of entire population from which the sample is drawn.

**Secondary data analysis**—A re-analysis of data using the same or other appropriate procedures to verify the accuracy of the results of the initial analysis or for answering different questions.

**Self-administered instrument**—A questionnaire or report completed by a study participant without the assistance of an interviewer.

**Significance**—Overall significance represents the total synthesis of all that has been learned about the merit or worth of the program or project. This is different from statistical significance, which may be testing one of several conditions of a program or project.

**Stakeholders**—Individuals and groups (both internal and external) who have an interest in the evaluation; that is, they are involved in or affected by the evaluation. Stakeholders may include program staff or volunteers, program participants, other community members, decisionmakers, and funding agencies.

**Standardized instruments**—Assessments, inventories, questionnaires, or interview protocols that have been tested with a large number of individuals and are designed to be administered to program participants in consistent manner. Results of tests with program participants can be compared to reported results of the tests used with other populations.

**Standardized tests**—Tests that have standardized instructions for administration, use, scoring, and interpretation, with standard printed forms and content. They are usually norm-referenced tests but can also be criterion-referenced tests.

**Statistic**—A summary number that is typically used to describe a characteristic of a sample.

**Statistical procedures**—The set of standards and rules based in statistical theory, by which one can describe and evaluate what has occurred.

**Statistical test**—Type of statistical procedure, such as a *t*-test or Z-score, that is applied to data to determine whether your results are statistically significant (i.e., the outcome is not likely to have resulted by chance alone).

**Structured interview**—An interview in which the interviewer asks questions from a detailed guide that contains the questions to be asked and the specific areas for probing.

**Summary**—A short restatement of the main points of a report.

**Summative (or outcome or impact) evaluation**—A type of evaluation that assesses the results or outcomes of a program. This type of evaluation is concerned with a program's overall

effectiveness. It presents conclusions about the merit or worth of an intervention and recommendations about whether it should be retained, altered, or eliminated.

**Time series study**—A study in which periodic measurements are obtained prior to, during, and following the introduction of an intervention or treatment in order to reach conclusions about the effect of the intervention.

**Treatment**—Whatever is being investigated; in particular, whatever is being applied to, supplied to, or done by the experimental group that is intended to distinguish them from the comparison groups.

**Treatment (or experimental) group**—A group of individuals receiving the treatment or intervention being evaluated or studied.

**Triangulation**—In an evaluation, triangulation is an attempt to get a fix on a phenomenon or measurement by approaching it via several (three or more) independent routes. For example, it might involve obtaining data on the same variable from two or more sources.

**Unanticipated outcomes**—An unexpected result of a program or treatment.

**Utility**—The extent to which an evaluation produces and disseminates reports that inform relevant audiences and have beneficial impact on their work.

**Utilization** (of evaluations)—The extent to which evaluation results are used to inform decisions or actions.

**Validity**—The extent to which a measurement instrument or test accurately measures what it is supposed to measure. For example, a reading test is a valid measure of reading skills, but is not a valid measure of total language competency.

**Variables**—Specific characteristics or attributes, such as behaviors, age, or test scores, that are expected to change or vary.