DOCUMENT RESUME

ED 445 403	EA 030 609

AUTHOR	Yap, Kim; Aldersebaes, Inge; Railsback, Jennifer; Shaughnessy, Joan; Speth, Timothy
TITLE	Evaluating Whole-School Reform Efforts: A Guide for District and School Staff. Second Edition.
INSTITUTION	Northwest Regional Educational Lab., Portland, OR.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	2000-08-00
NOTE	173p.
CONTRACT	S283A50041-99C
PUB TYPE	Guides - Non-Classroom (055)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Educational Administration; Educational Change; Elementary Secondary Education; *Formative Evaluation; *Program Evaluation; School Organization; School Restructuring
IDENTIFIERS	Comprehensive School Reform Demonstration Program

ABSTRACT

This guidebook provides evaluation assistance to district and school staff. It was published in response to the Comprehensive School Reform Demonstration (CSRD) Program, passed by Congress in 1997 to provide incentives and support for low-performing, high-poverty schools. CSRD is an attempt to ensure that schools conduct evaluation of whole-school reform efforts in a way that provides valid and useful information for accountability and program improvement. The guide does not examine the philosophical underpinnings of evaluation issues. Rather, it provides guideposts that district and school staff can consider in choosing an approach to evaluating their school-reform efforts. It is intended for use by school staff at sites that have stated goals for student achievement and have already decided on one or more comprehensive strategies for reaching their goals. The text is arranged in a "train the trainer" format and is organized so as to assist in the design of a professional-development workshop. The book focuses on implementation evaluation and impact evaluation. Various design samples are also included to help schools customize their evaluation efforts. Each chapter includes handouts, small-group activities, transparency masters, and step-by-step instructions for creating an effective evaluation. A list of print and online resources appears in the back. (RJM)

Evaluating Whole-School Reform Efforts

A Guide for District and School Staff

Second Edition
August 2000

Northwest Regional Educational Laboratory

The contents of this publication were developed under Grant No. S283A50041-99C from the U.S. Department of Education. However, the contents do not necessarily represent the policy of the Department of Education, and endorsement of the contents by the federal government should not be implied.

Evaluating Whole-School Reform Efforts

A Guide for District and School Staff

Second Edition
August 2000

Kim Yap
Inge Aldersebaes
Jennifer Railsback
Joan Shaughnessy
Timothy Speth

Comprehensive Center, Region X

Northwest Regional Educational Laboratory
101 SW Main, Suite 500 Portland, Oregon 97204

Table of Contents

Figures

Tables

Acknowledgments

This guidebook is the result of a collaborative effort among many individuals who are interested in the improvement of teaching and learning through comprehensive school reform.

At the U.S. Department of Education, Patricia Gore, Sharon Saez, and Joyce Murphy provided leadership and support that made the project possible.

At the Northwest Regional Educational Laboratory, Rex Hagans and Steve Nelson reviewed our original proposal for developing the guidebook and graciously provided staff support to ensure the successful completion of the project. Jane Grimstad assisted with the design of the guidebook.

A conference was held in Portland on September 24, 1999, to review a preliminary draft of the guidebook and to obtain feedback for revision and refinement. We wish to thank all conference participants, including the Comprehensive Center advisory committee members, who shared their ideas, observations, and insights that have enriched the content and enhanced the format of the guidebook.

A special note of thanks is extended to Jana Potter who designed the process for obtaining maximum input from participants at the conference. We are grateful to Georgia Pfleiderer and Audrey Trubshaw for their assistance with logistical arrangements for the conference.

Finally, we wish to thank Lee Sherman and Catherine Paglin for their superb editorial help and Marjorie Wolfe and Shannon Hillemeyer for their excellent assistance with word processing, design, and production of the guidebook.

Introduction

In the fall of 1997, Congress set in motion a federal initiative to jump-start comprehensive reform in the nation's schools. The Comprehensive School Reform Demonstration (CSRD) Program provided incentives and support for schools, particularly high-poverty, low-performing schools, to develop and implement comprehensive school reform efforts. These schools are to carry out reform activities based on reliable research and effective practice. In the fall of 1998, Congress extended CSRD funding for a second year.

The current Title I legislation also provides an incentive for schools serving a high concentration of poor children to engage in whole-school reform. Title I schoolwide programs, implemented in schools with at least 50 percent of students in poverty, have the flexibility of pooling resources from other federal programs to plan and implement schoolwide improvement activities.

The most recent U.S. Department of Education estimate indicates that there are 1,600 schools participating in the Comprehensive School Reform Demonstration (CSRD) Program across the nation. In addition, approximately 15,000 Title I schools are implementing schoolwide programs. Each of these whole-school reform efforts is to be evaluated to assess its impact on teaching and learning.

The Education Department has issued general guidance to help district and school staff evaluate CSRD and Title I schoolwide programs. This guidebook, developed collaboratively by the Comprehensive Center and the CSRD work unit at the Northwest Regional Educational Laboratory, is intended to provide further evaluation assistance to district and school staff. It is an attempt to help ensure that schools conduct evaluation of whole-school reform efforts in a way that provides valid and useful information for accountability and program improvement.

The guidebook is not intended to be a philosophical discussion of evaluation issues. Nor is it designed to be a cookbook on the evaluation of whole-school reform efforts. Users who have no prior training or experience with program evaluation will not become skilled evaluators by reading the document. Rather, it is our intention to provide some guideposts that district and school staff can consider in choosing an approach to evaluating their school reform efforts. We hope that this guidebook will help raise awareness of the complexity of program evaluation in general and the evaluation of whole-school reform efforts in particular.

Overview

The Purpose of This Guide

The intention of this guidebook is to increase understanding about how to design and implement an evaluation plan that will help answer questions about program quality and effectiveness in accomplishing school improvement goals. Rather than turning to outside sources for evaluation expertise, schools can build their own knowledge and skills about how to evaluate whole-school reform efforts. As a result, schools will gain confidence in their ability to demonstrate that their efforts are making a difference in student achievement, as well as meet growing accountability requirements.

This guide is to be used by school staff at sites that have already specified goals for student achievement (as required in most grant applications), and have also decided on one or more comprehensive strategies for reaching their goals. Once this preliminary planning work has been done, the school will be in a position to draw upon the information presented in this guidebook to develop a useful evaluation plan.

The guidebook has been planned to assist in the design of a professional development workshop. It is arranged in a "train the trainer" format. The hope is that those responsible for evaluation will use this guide to provide staff development for all individuals who are engaged in comprehensive school reform, with the purpose of increasing their knowledge and involvement in the evaluation process.

A wealth of information and activities is organized into specific sections that can be presented together or separately, depending on the needs of the workshop audience. Workshop audiences can vary; possible participants include an entire school staff, a leadership team that is responsible for the implementation of a CSRD Program, or Title I schoolwide school principals within a district. Each section furnishes the presenter with an explanation of various aspects of evaluation design and process, instructions for carrying out the workshop, and corresponding activities and transparencies.

Who Is Responsible for Evaluation?

Often, school staff ask, "Who should be the evaluator?" The greatest benefits from evaluation are realized when the school takes ownership of evaluation and uses the findings to stimulate change that makes a difference in how they go about comprehensive reform. For this reason, the best answer to this question is an evaluation team composed of representatives from the whole-school community. We highly encourage schools to include any group or person that has an investment in either the implementation or results of its school reform efforts.

An internal evaluation team will increase the likelihood that the evaluation plan will be administered well. The evaluation team's responsibility begins with designing a relevant evaluation plan that addresses their information needs and grant requirements. This requires generating enthusiasm and support in the school community for the evaluation plan. A significant role the evaluation team will be assigned

Glossary

School Community–all individuals and groups who have an invested interest in the school, for example, students, parents, teachers, local employers, principals, or school board members

Program Evaluation–the use of various methods to determine the degree to which a program has been developed and implemented as planned, as well as accomplished its stated goals and objectives

Formative Evaluation–the monitoring of activities and strategies that take place during the development and implementation of a program and informs stakeholders about possible program adjustments to improve quality and effectiveness

Summative Evaluation–evaluation of the ultimate results of a program, asking the question, "Has the program accomplished what it intended?"

is the administration of the evaluation plan, which involves identifying and developing instruments, collecting necessary data, analyzing and interpreting data, and reporting results to all stakeholders. Dedicating sufficient time and resources is essential to the success of an evaluation plan. The evaluation team will need time to oversee the evaluation process, immerse themselves in the data, and ensure that findings are considered throughout program implementation and converted into constructive changes that improve school improvement efforts.

Another consideration for evaluation is when to use external expertise, an outside evaluator. Outside evaluators can provide the technical guidance during the design, analysis, and reporting phases of evaluation. The evaluator's assistance will help ensure that evaluation plans are relevant and realistic. The collection of data is typically left to the project staff because of the ease and day-to-day access they generally have to data. A collaborative working relationship between the outside evaluator and internal evaluation team merges the best of two viewpoints. An outside evaluator brings an objective perspective to the process and can more easily ask the difficult, reflective questions that can be missed/avoided by those who are implementing the comprehensive reform program. Further, the project staff comprehend the data best and can attach meaning to the numbers generated by an evaluation. By collaborating with an outside evaluator, schools can overcome some of the typical obstacles (fear, lack of experience in evaluation, time limitations) they face when planning and implementing evaluation plans, thus increasing the feasibility of their evaluation strategies.

Evaluation Requirements

Specific evaluation requirements for state or federal grants have been purposely left out of this guide. We have chosen not to address such requirement issues because often the requirements are explicit to a grant, differ from year to year, and vary from program to program as well as from state to state. For these reasons, it would be very difficult to accurately address requirement issues around evaluation. The best approach to ensure that your school's evaluation plan meets program/grant specific requirements is to contact your state's educational agency.

Context of Comprehensive School Reform

Comprehensive school reform and Title I schoolwide programs are well underway across the nation. Along with being responsible for restructuring their operational systems, schools increasingly are being held accountable for the results of their whole-school reform efforts. Federal and state education officials are asking several significant questions: (1) Are comprehensive school reform efforts producing positive results in student achievement? (2) Are comprehensive school reform programs being implemented as planned and with fidelity to the adopted model? and (3) Will state and local policies and practices sustain comprehensive school reform? These questions should drive evaluation efforts. (Overview Transparency #1)

Glossary

Outcome–immediate effects or results of a program

Impact–long-term effects or results of a program

Performance Indicator–measures designed to provide data to signify the extent to which a specific program objective is achieved

The overarching goals of evaluation are twofold: to inform schools about what is and isn't working, and to guide decisions about program adjustments and improvements, thereby increasing the likelihood of positive impact.

Program evaluation is a systematic process designed to gauge the quality and effectiveness of a program. Evaluation produces information that helps monitor progress and solves problems to enhance program implementation and impact. Evaluation is most meaningful when it is integrated early into the program design. Tacking it at the end of a program seldom yields useful findings. (Overview Transparency #2)

There are two basic types of evaluation, each with its distinct purpose. "Formative" evaluation produces information used to improve a program during its operation. It generates information that guides decisionmaking about the program's desirability, feasibility, fidelity, and soundness in producing desired results (Nelson, 1999; Sarvela & McDermott, 1993). "Summative" evaluation, on the other hand, garners data necessary for judging the ultimate success of the entire program (Sarvela & McDermott, 1993). Its major purpose is to answer the question, "Did the program do what it promised?" (Overview Transparency #4)

Often, evaluation focuses only on results. But without data on program implementation, it is difficult to link student outcomes to the program or to make timely adjustments to enhance program effectiveness. With ongoing and well-thought-out program evaluation, a school community can construct a compelling case that its comprehensive reform efforts did indeed contribute to the improvement of its students' academic performance.

A number of assumptions guide program evaluation (Northwest Regional Educational Laboratory [NWREL], 2000). It should:

- Be comprehensive enough to reflect decisionmaking needs and provide timelines for ongoing, immediate feedback for continuous program improvement
- Use a multimethod approach to enhance the validity of data
- Provide sound information regarding outcomes and effectiveness in achieving expected program outcomes
- Employ a combination of quantitative and qualitative strategies

Program Implementation

Research has consistently shown that the depth and quality of program implementation is a powerful factor in the success of school reform programs. Comprehensive reform efforts can succeed if they are implemented well. In particular, schools should pay attention to how widely staff members embrace the program and how well they understand it. Schools should ask, "Is the program being implemented as intended?" Research has identified nine program components (see sidebar on page 6) that contribute to the quality of a comprehensive reform program and are influential in helping improve student achievement. Careful monitoring of these nine components provides insight into what factors help or hinder reform efforts. These components can provide a useful framework for gathering, interpreting, and using data to make decisions about implementation progress and challenges. The specific evaluation questions that guide the process and determine which data collection strategies to use are (Sarvela & McDermott, 1993): (1) Which intervention activities are being used? (2) Is the intervention being implemented with fidelity? (3) What is working? (4) What should be improved? and (5) How should it be refined? Answers to these questions help determine how a school's reform program is making a difference. Linking achievements to comprehensive school reform efforts is then possible. (Turn to Page 19 for detailed discussion.)

Program Outcome and Impact

Summative evaluation involves gathering the evidence necessary to determine overall program success in improving student achievement. The evaluation question driving this portion of the investigation is, "Are we achieving what we aspired to do?" In the context of comprehensive school reform, program success is measured by how well the school stacks up against state standards and local assessment measures.

Many reasons and benefits warrant conducting program evaluations, including (Overview Transparency #3):

- Strengthen program design by clearly articulating shared goals and objectives
- Facilitating informed decisionmaking about improving the quality of the program
- Contributing to making constructive changes to enhance program effectiveness
- Helping identify and celebrate successes when desired outcomes are achieved
- Reinforcing the link between schoolwide program strategies and student outcomes

There are two basic forms of summative evaluation: outcome evaluation and impact evaluation. (Overview Transparency #6) Outcome evaluation examines immediate changes in knowledge, skill, attitude, and behavior. Impact evaluation, on the other hand, demonstrates the program's long-term effects (Muraskin, 1993). Here's an example: A school gives a parent workshop about the value of reading to children at home. The program outcome would be the new knowledge parents gained from their participation. This direct effect—increased parental knowledge—is an immediate result that may lead to increased reading with children at home. This in turn leads to a positive impact on academic achievement. In the world of evaluation, both new parental knowledge and more reading at home would be considered program outcomes. Improved reading achievement would be considered a long-term program impact.

Routinely monitoring outcomes is beneficial because it provides frequent feedback to those involved in decisionmaking about the program. Knowledge gained from monitoring outcomes can gauge progress, uncover problems, help appropriately allocate resources, and acknowledge successes (Pane, Mulligan, Ginsburg, & Lauland, 1999). For example, if a program objective is to increase reading scores on the state assessment by 10 percent over the next three years, outcomes (such as improved reading skills) will help determine whether the school is moving in the desired direction. Program outcomes are results that are related to an objective but that occur more immediately. Knowing precisely what outcomes the school is looking for will help ascertain which data sources contain the desired information. This can help schools avoid the common error of collecting unneeded data that can hike costs and waste time. (Turn to Page 49 for detailed discussion.)

These components, when integrated into comprehensive school reform plans, enhance the quality and effectiveness of a program (Overview Transparency #5):

- **Innovative strategies** and proven methods that are based on reliable research and replicated successfully in schools with diverse characteristics
- A **comprehensive design** for effective school functioning
- **Measurable goals** for student performance and benchmarks for meeting those goals
- Commitment and **support of school staff** and community
- **Meaningful involvement** of parents and local community
- High-quality **external technical support** and assistance
- **Evaluation plan** for monitoring program implementation and assessing results in student achievement
- **Coordinated resources** to maximize and sustain the school reform effort
- High-quality and continuous teacher and staff **professional development**

Evaluation Design and Process

How does a school design a comprehensive evaluation plan that meets federal and state requirements, and also satisfies its own informational needs? By addressing certain key questions early in program planning, the evaluation process will reflect the needs, interests, issues, and resources unique to the school (Sarvela & McDermott, 1993; Western Regional Center, 1995). Questions that schools should ask of themselves are (Overview Transparency #7):

- *What does our school want to accomplish overall?*

This requires clearly articulating goals and transforming them into specific, measurable objectives. Setting goals and objectives is difficult. Your school must first consider current conditions, needs, academic concerns, and resources. Creating a snapshot of your school can help you avoid the common pitfall of setting goals and objectives that are unrealistic given the available resources. The value of conducting a thorough needs assessment cannot be overemphasized. It will clarify issues, pinpoint priorities, and identify resources.

- *What will our school have to do to achieve these goals and objectives?*

This is the stage when your school decides on specific strategies and activities to create the desired changes. This is when you determine how program goals and objectives are translated into research-based actions and strategies. Actions and strategies should match goals and needs. Without that match, your school will have a tough time reaching its objectives.

■ *How will our school know that its program is succeeding at accomplishing its goals and objectives?*

Schools can gauge progress toward their goals by selecting program and student performance measures that are meaningful, measurable, and relevant—that is, related to program objectives. Performance indicators will provide the information needed to demonstrate program success. It's best to measure progress annually and at interim checkpoints (say, quarterly). With regular monitoring, your school can uncover barriers to success and devise new strategies as you go along.

■ *How will evidence be gathered to demonstrate progress toward our school's goals?*

At this point schools need to decide which data collection methods they will use to acquire relevant information. (Turn to Page 26 for detailed discussion.) Typically, schools have a wealth of information at hand because they are continually gathering data for various purposes. For this reason, schools can begin by building on existing systems, adding only data collection methods that will fill information gaps. Data collection methods are many. They include document review, surveys, interviews, focus groups, observation, and student achievement assessments. Ideally, schools will choose to use multiple data gathering procedures to improve the credibility of their data. For example, changes in teaching practices can be assessed in several ways: administering a survey to students, observing classroom practices, or conducting a focus group with teachers. Using two or three data collection methods, measurement instruments, or data sources is a technique called "triangulation." Each data gathering method has advantages and disadvantages. (Turn to Page 67 for Data Collection Matrix.)

Glossary

Triangulation–confirming data credibility by using multiple data-gathering methods or multiple sources of data.

Disaggregation of data–comparing of subgroups based on demographic characteristics and educational experiences that are deemed important.

■ *How will our school determine what the data are telling us?*

Making sense of the data collected becomes essential if the findings are to be used to influence decisions and future planning about the school's comprehensive reform efforts. Interpretation of the data is best accomplished when it is reviewed by the school's staff and community, in particular those who are responsible for the day-to-day implementation of the program. Data analysis is an inquiry process meant to help schools examine and better understand the nature and effectiveness of their school improvement program. The following are reflective questions that can help guide discussions (Holcomb, 1999; Levesque, Bradley, Rossi, & Teitelbaum 1998): (1) What do these data reveal? (2) What else might explain these results? (3) What else do we need to know to better understand the data before we draw conclusions? (4) What good news is here for us to celebrate? and (5) What needs to be done to improve program performance and effectiveness?

■ *How will our school use evaluation results?*

To maximize the benefits of evaluation, schools should establish an ongoing process to review, interpret, and communicate results. In this way, schools can keep the school community informed about the program's quality and effectiveness. Sharing successes generates enthusiasm, involvement, and commitment to the reform program.

The same people who are implementing the program should collect and interpret the data. In this way, they will get immediate feedback to inform daily decisions about program operations and classroom practices. Besides getting ongoing feedback, the school staff and community gain a sense of ownership by direct involvement. Ownership develops intrinsic motivation to carry out the evaluation plans, interpret results, draw conclusions about program progress, and pursue improvements. Most of all, it fosters trust that data will be used in a positive, not punitive, way.

Common Barriers to the Collection and Use of Evaluation Data

- Challenge of collaboration
- Lack of time
- Lack of proper training in practical program evaluation
- Fear of evaluation

The school's evaluation plan arises from thoughtful consideration of these questions. Well-designed evaluations are invisible, becoming imbedded in daily routines. The most useful evaluation plans are those that are tailored to the unique needs and context of the reform program. The best plans glean relevant information about program performance and student achievement that will contribute to maximizing the program's effectiveness.

To make sure their evaluation plan succeeds, schools must address the reasons people resist evaluation. Common barriers to the collection and use of evaluation data include:

- *Challenge of collaboration.* Staff, parents, and administrators often lack not only sufficient time to work collaboratively but also the skills and experience to work cooperatively.
- *Lack of time.* The most common obstacle is the shortage of time to successfully plan and implement evaluation. Many teachers already feel overwhelmed, and the thought of one more thing to do can be daunting.
- *Lack of proper training in practical program evaluation.* Few have the knowledge, skills, or confidence to conduct program evaluations or the understanding of how to use data to guide decisions.
- *Fear of evaluation.* Many educators fear that data will be used against schools by exposing inadequacies and jeopardizing funding. This fear stems mainly from a misperception about the purpose and function of evaluation.

Use of Data for Program Improvement

Evaluation is meaningless unless data are collected, reviewed, analyzed, and disseminated quickly and efficiently. Only when results are fed back into the system are they useful. The process of interpreting and reporting evaluation results is most meaningful when it is part of an ongoing, evolving process that engages all interested people. Schools must invest time to review and interpret results in order to realize the benefits of evaluation.

Whenever possible, data should be disaggregated—that is, broken down by categories such as gender, ethnicity, student type, and grade level. By disaggregating data, schools can zero in on areas of strength and weakness. Disaggregation of data also helps schools better understand the program's impact, in addition to addressing equity issues (Yap, 1997).

Strengthening Programs Through Evaluation

Evaluation is a powerful tool that can reveal what is actually occurring in schools. It can sift through the maze of school reform efforts to uncover what is truly working to change the learning environment. It can reveal the root causes of schools' struggles so that the real problem—not just the symptoms—can be tackled. It can also bring to light factors that contribute to positive results so that schools can continue to improve teaching and learning.

No strand of a school—from curriculum and instruction to facilities operation, staff development, and administration—goes untouched in the schoolwide reform process. The goal is to deliver a coherent, sound education that will bring high standards within reach for each and every child. Evaluation is the means of finding out where your school has been, where it's going, how it's getting there, and—most important—whether it's on target to reach its desired destination. If goals and practices are out of sync, evaluation can point the way to get back on track.

In the following sections of this guide, your school community will find the step-by-step guidance it needs to plan, design, and carry out effective evaluation of your comprehensive school reform program.

References

Ary, D., Jacobs, L.C., & Razavieh, A. (1996). *Introduction to research in education* (5th ed.). Fort Worth, TX: Harcourt Brace College.

Holcomb, E.L. (1999). *Getting excited about data: How to combine people, passion, and proof.* Thousand Oaks, CA: Sage.

Levesque, K., Bradby, D., Rossi, K., & Teitelbaum, P. (1998). *At your fingertips: Using everyday data to improve schools.* Berkeley, CA: MPR Associates, Arlington, VA: American Association of School Administrators, & Berkeley, CA: National Center for Research in Vocational Education.

Muraskin, L. (1993). *Understanding evaluation: The way to better prevention programs.* Rockville, MD: Westat.

Nelson, S. (1999, June). *Principles of evaluation.* Paper presented at Charter Schools Leadership Training Academy, Portland, OR.

Northwest Regional Educational Laboratory. (2000). *Developing your school's CSRD evaluation plan: An awareness workshop for local schools [Training materials].* Portland, OR: Author.

Pane, N., Mulligan, I., Ginsburg, A., & Lauland, A. (1999). *A guide to continuous improvement management (CIM): For 21st century community learning centers.* Washington, DC: U.S. Department of Education.

Sarvela, P.D., & McDermott, R.J. (1993). *Health education evaluation and measurement: A practitioner's perspective.* Madison, WI: Brown & Benchmark.

Western Regional Center for Drug-Free Schools and Communities. (1995). *Systemic evaluation: A new approach to assessing the effects of tobacco, alcohol, and other drug (TAOD) programs.* Portland, OR: Northwest Regional Educational Laboratory.

Yap, K.O. (1997). *Guidebook on developing performance indicators.* Portland, OR: Northwest Regional Educational Laboratory.

Instructions for Overview Transparencies

Each transparency is related to the Overview section of the guidebook. Becoming familiar with the contents of this section will help guide your use of the transparencies. This section of the guidebook and corresponding transparencies provide a conceptual overview with brief description of critical elements of program evaluation. More indepth discussions and examples of how to design and plan for program evaluation will be presented later in the guidebook.

Transparency #1

Sets the stage for understanding the significant overall questions driving comprehensive school reform evaluation. Briefly discuss as described on Page 4 in the guidebook.

Transparency #2

Describes the overall purpose of program evaluation. Distinguishing between the two dimensions of formative (implementation) and summative (impact) evaluation is useful in helping understand the unique purpose of each. Briefly discuss as described on Pages 4-5 in the guidebook.

Transparency #3

Outlines the benefits of evaluation with particular attention to its value in guiding decisions to improve the effectiveness of the comprehensive reform program. Briefly discuss as described on Pages 5-6 in the guidebook.

Transparency #4

Provides a brief comparison of formative and summative evaluation purpose and data collection methods.

Transparency #5

Discusses nine components of effective comprehensive school reform. Briefly discuss as described on Page 6 in the guidebook.

Transparency #6

Introduces program outcome and impact evaluation. Briefly discuss as described on Pages 5-6 in the guidebook.

Transparency #7

Introduces questions that facilitate the planning of program evaluation. Briefly discuss as described on Pages 6-7 in the guidebook.

Transparency #8

Introduces the common barriers that often confront schools when planning and implementing evaluation plans. Briefly discuss as described on Page 8 in the guidebook.

The Significant Questions Driving Evaluation

1. Are comprehensive school reform efforts producing positive results in student achievement?

2. Are comprehensive school reform programs being implemented as planned and with fidelity to the adopted model?

3. How will state and local policies and practices help sustain comprehensive school reform?

Program Evaluation

- Program evaluation is a systematic process that is designed to determine the quality and effectiveness of a particular program

- Formative evaluation generates information used to guide decisionmaking about the program's desirability, feasibility, fidelity, and soundness in producing desired results

- Summative evaluation involves the collection of data necessary for judging the ultimate success of the entire program

Benefits of Evaluation

- Strengthens program design by clearly articulating shared goals and objectives
- Facilitates informed decisionmaking about improving the quality of the program
- Contributes to making constructive changes to enhance program effectiveness
- Helps identify and celebrate successes when desired outcomes are achieved
- Reinforces the link between schoolwide program strategies and student outcomes

Comparisons of Formative and Summative Evaluation

Issue	Formative Evaluation	Summative Evaluation
Purpose	Program improvement – ongoing process of providing feedback about the quality and effectiveness of the program	Program achievement – process of assessing the degree to which the program has accomplished predetermined goals
Overall Evaluation Questions	Is the program being implemented as planned to accomplish its intended goals?	Are the efforts having a positive impact on student achievement?
Specific Evaluation Questions	What is working? What should be improved? How should it be changed? Which interventions are being used? Is the intervention being implemented with fidelity?	What has happened? Who was affected? What was the most effective treatment? Was it cost-effective?
Measurement Methods	Surveys Interviews Observation Review documents and artifacts	State standards and benchmarks Local student achievement assessments - Standardized tests - Local assessments (IRIs, writing assessments, math) Student performance records - Graduation, dropout rates - Attendance, suspensions

22

23

Reference: Health Education Evaluation and Measurement – Paul Sarvela and Robert McDermott

Nine Comprehensive Components

1. Innovative strategies based on reliable research

2. Comprehensive design for effective school functioning

3. Measurable goals and objectives

4. Support of school staff

5. Meaningful involvement of parents and local community

6. External technical support and assistance

7. Evaluation plan

8. Coordinated resources

9. Ongoing professional development

Program Outcome and Impact

Outcome—examines immediate changes in knowledge, skills, attitudes, and behavior.

For example: *Increased parent knowledge about the value of reading at home with their children as a result of a parent workshop.*

Impact—the long-term effects of the intervention.

For example: *Improved reading scores on ITBS.*

Key Questions That Shape the Evaluation Planning Process

- What does your school want to accomplish overall?
- What will your school have to do to achieve these goals and objectives?
- How will your school know that its program is succeeding at accomplishing its goals and objectives?
- How will evidence be gathered to demonstrate progress toward your school's goals?
- How will our school determine what the data are telling us?
- How will your school use evaluation results?

Common Issues and Challenges

- Challenge of collaboration
- Lack of time and resources
- Lack of proper training in "practical" program evaluation
- Fear of evaluation

Implementation Evaluation

Let's imagine that your school has chosen a new comprehensive or schoolwide program. If your school is seriously committed to getting this program in place, then the implementation efforts cannot be left to chance. A process for verifying progress will have to be an integral part of the work. A critical initial step is to plan an evaluation that will transport detailed information about program implementation back to the program planners. This type of evaluation—collecting and using data to feed back into the program on an ongoing basis—is called formative evaluation.

Formative evaluation serves two purposes:

1. To determine whether the program is being implemented as the program developers designed it and that the most vital components of the program are in place

2. To enable staff to retool and fine-tune their efforts to make a program work at a specific site

A strong formative evaluation can help a program to "hum" at a particular school.

The central question in formative evaluation is whether the model or program is being implemented as it was designed. Comprehensive models are grounded in research. But no program—no matter how sound it is—can have impact if its essential elements are not used. If some staff choose to use only a portion of a new program and to selectively abandon other parts of the program, they weaken the impact of that program. This is why systematic data collection about implementation is needed. By determining which program components are firmly in place and which ones are only being given lip service, those managing the new program can learn about and address the barriers that are limiting or interfering with use. They can also design special adaptations to meet specific needs of this school.

In implementation evaluation, the data collected are used primarily for internal reporting to the program staff (although some grants do require that implementation data be reported to the funding agency[1]). To maximize the potential for program improvement, evaluation data about implementation must be analyzed quickly, shared broadly, and presented in a format that can be easily used to make program modifications. Implementation evaluation works best when the evaluation is seen as an integral part of staff development.

An important decision is to identify a team of individuals who can collect the implementation evaluation data. The evaluation of a comprehensive program is best done by a team of data collectors. This team could include external evaluators, the administrators or staff in the building, and parents and community members. To be effective, members of that team need to be able to meet regularly with those implementing the program so there are clear lines of communication and a thorough understanding of the evaluation work. Step-by-step guidance for this team is presented in the following pages. Before getting into that level of detail, it is important to reiterate the key elements that research has shown to predict successful implementation. Throughout this data collection, *all* involved in program implementation need to be aware of these factors so they can gather evidence to verify that the necessary supporting conditions exist and that specific instructional components are making it into the classroom to bolster the comprehensive reform.

Formative or implementation evaluation is designed to provide data that will refine and improve a program. The purpose of doing such an evaluation is to gather adequate data to ensure that a program works in the local context.

[1] Prior to developing an evaluation plan, program managers need to review the evaluation requirements stipulated by their funding agency. They should also determine how the data provided in reports will be used. Program managers need to be very clear about whether decisions about continued funding will be made based upon the reports they submit.

Research shows that to be effective, comprehensive reform needs to:

- Be undertaken for the right reasons (for example, to solve a problem, meet a need, or improve student achievement), not simply to advance the career of an administrator or to procure additional funds.
- Nurture commitment on the part of teachers, preferably by involving them from the beginning in discussions of what and how to change.
- Provide adequate resources, including funds, materials, and—most important—time for teachers to learn, practice, reflect, discuss, observe, evaluate, and assimilate.
- Include ongoing professional development for teachers, not depend on a one-shot training workshop at the beginning of implementation. Training and coaching should be ongoing and should support the change of classroom practice.
- Promote collaboration among teachers so they can learn from each other and help each other work through the most difficult aspects of change.
- Exert pressure on teachers who are resistant to change and develop approaches that channel resistance into productive dialogue. To prevent resentment and passive resistance, this pressure must be counterbalanced by continuous support.
- Enable staff to try new and messy changes by allowing them to make mistakes and encouraging them to make midcourse adjustments.
- Involve parents and community members in the reform process.
- Ensure that school and district leaders support the change in word and deed.
- Minimize conflicts with other innovations, programs, and policies.
- Incorporate successful innovations into district policy and budgets so that they will outlast the inevitable departure of key leaders or start-up funding (Buechler, 1997).

One aspect of the implementation evaluation is to determine if these basic conditions are being met.

Considerations in Planning Your Implementation Evaluation

The degree and depth of implementation evaluation a school is able to undertake depends on two pragmatic factors: amount of funding and access to data. Hiring an external evaluator is an excellent way to get this work done but since implementation evaluation can be very time intensive, contracting with an outside consultant to do all data collection work may be more costly than most projects can afford. In addition, physical distance from the day-to-day operations may restrict the amount of detailed information that an external evaluator can collect. For these reasons, many programs use a combination of external and internal staff to collect data.

In planning this evaluation it is also essential for the program managers to carefully study grant requirements to determine the type of evaluation data which is required and to know how this data will be used. Most funders require that outcome data be reported, so it is easy for a school to be focused exclusively on this type of data. Schools must be careful not to become focused exclusively on end results, to the detriment of ongoing measures of implementation. Implementation measures are critical to achieving the long-term results schools seek.

Priorities of the preparatory work are:

- Addressing staff misunderstandings about program evaluation
- Getting staff connected to the evaluation work so that they can participate in question generation and data collection

Preparation for Evaluation

This section describes concrete ways that school staff can get more engaged in the process of posing evaluation questions and identifying how data will affect program implementation. These steps will help guide the initial work of the implementation evaluation:

Step 1: Orient the entire staff to evaluation issues as early as possible. The primary source of information for implementation evaluation is likely to be front-line staff—those who are working to put this program into place. Since these individuals will be supplying information, it is crucial that they understand the purpose of evaluation and are willing to help collect data. Those collecting the evaluation data need to make sure that all participants have been informed about the purpose of the evaluation and are willing to be cooperative. If those collecting evaluation data already have the trust of the staff, they may want to proceed to Step 2. If not, we suggest preparatory work (described below) to ensure that staff are able and willing to cooperate fully and provide the best information possible about program implementation.

Those conducting the evaluation need to address any misunderstandings or reservations staff may hold about the process of evaluation. When personal concerns about evaluation have been discussed, staff will be more willing to provide honest data. The following issues often crop up:[2]

- Staff equate *program* evaluation with *personnel* evaluation. When a program is being evaluated, staff can take this very personally. They may feel that it is they who are being critiqued and this puts them on the defensive. One way to address this is to explain the difference between studying individual performance and examining the complex system in which a program operates. Individuals operating alone in a complex system benefit from a better understanding of how systemwide change happens. The evaluator can help the staff understand that for any program to work, all people involved need to get beyond assigning blame and join together to address the big issues.

Several issues about the evaluation process should be clarified with program staff. The intention is to raise staff awareness of the usefulness and power of the evaluation work being done at the school.

- Some staff believe that evaluation data will be used exclusively to decide if a program will be refunded. Naturally, they are reluctant to reveal any problems, concerns, or weaknesses if they think that making such information public will mean the elimination of program funds. Staff need to understand that the purpose of implementation evaluation is program improvement, not funding decisionmaking. Being clear about how specific information will be used is essential.
- Staff may believe that evaluation needs to be done by an impartial observer. They may think they should keep their distance to avoid "contaminating" the data. The evaluator needs to stress the importance of staff involvement and participation in evaluation.
- If external evaluators participate in this data collection, they need to clarify their own role and function. Those collecting the data need to explain that they envision the evaluation process as a way to learn, rather than as a chance to criticize.

Research has shown that staff cooperation and understanding help a school use formative evaluation to improve its implementation of comprehensive reform. To make that happen, the following points about evaluation need to be explained at staff meetings:[3]

1. Evaluation should be planned early. The earlier the data are collected, the more likely those data can be used during the course of the program.

2. The evaluation must include multiple perspectives such as ideas from school staff, from the district offices, and from the reform model trainers working with the school.

3. This program does not operate in isolation from the larger context of the school. To ensure that the evaluation tackles the background or contextual issues, the evaluation process needs to examine the supportiveness of school culture and district policies for schoolwide reform. Staff should be aware that evaluation work may include reviews of other programs in the school to see where and how multiple programs overlap.

4. The evaluation will be looking at how staff development is incorporated into classroom instruction and management. This will mean classroom visits to monitor and assess program

[2] A summary of these issues is available in Implementation Transparency #1.

[3] These ideas are summarized in Implementation Transparency #2.

implementation. Those working on the evaluation should reassure staff that data about the work of individual teachers will be kept confidential. They should also stress that different rates of implementation across classrooms are natural.

5. Feedback from data collected will be provided to school staff as quickly as possible.

6. The same data will be collected repeatedly so that the school can assess progress. This means that when the school selects a data collection tool, it is making a commitment to use that instrument several times—either during the school year or for several years in a row. With this in mind, instrument selection needs to be done carefully and thoughtfully.

7. Much can be learned when the school's progress is compared to other reform efforts or national norms. To make such comparisons, schools may need to use measures that have been used in other settings.

Step 2: Initiate data collection and promote ongoing dialogue about evaluation at staff meetings and all meetings with parents and community members. Introduce evaluation concepts at staff and community meetings and take this opportunity to collect attitude and belief data. Four ideas for doing this are provided in the "Presenter's Guide and Training Materials." These activities can be used to spark staff and community conversations about the reform model. They also help evaluation planners understand the overall context for program implementation. The questions outlined in each activity can be adapted to each site and used to collect formative evaluation evidence at the beginning of any implementation process.

The evaluator can demonstrate possible pitfalls in implementation by citing the results from the recent RAND evaluation (Bodilly, Keltner, Purnell, Reichardt, & Schuyler, 1998), which documented a number of schools' efforts at schoolwide reform. The study focused on the Cincinnati School District, where three different models were implemented and supported by the district. When teachers were surveyed at the end of Year One about their new program, it was clear that many teachers who were supposed to be implementing the model were still uncertain about the work they were doing.

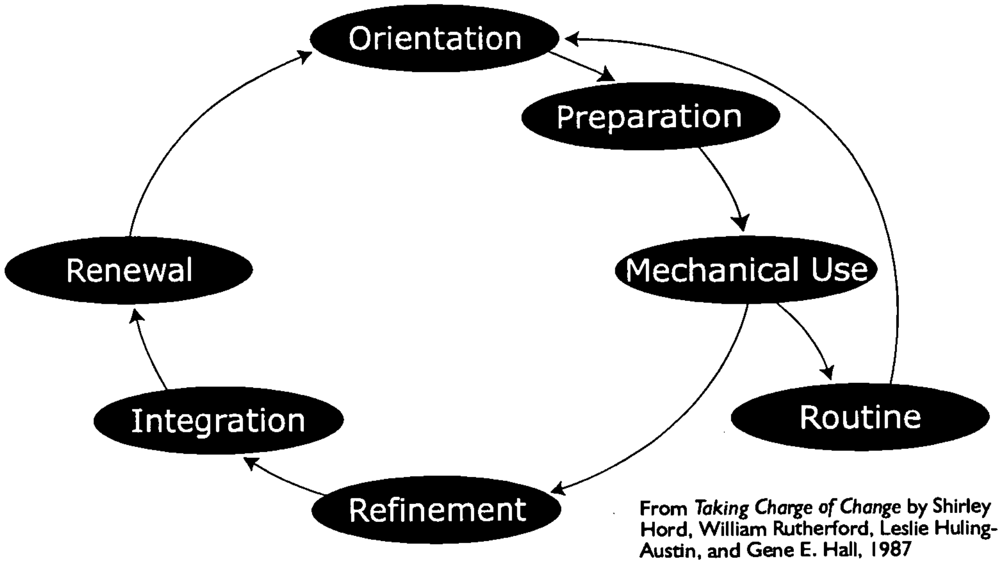
The RAND Study found that:

- ■ Only 57 percent of the teachers could identify which model was being used in their school
- ■ 27 percent felt they could explain the model's philosophy to others
- ■ 44 percent were unclear about success criteria (how their new program would be judged)
- ■ 38 percent felt that lack of success would lead to termination of the program
- ■ 22 percent felt that their personal efforts would affect the success of the design
- ■ 23 percent said they had strayed from the design (within certain designs, this was as high as 53 percent)

If the school district had known about teachers' lack of knowledge earlier in the year, it would have been able to remedy some of these implementation issues. This is where implementation evaluation can be helpful.

Step 3: Discuss the way program implementation is likely to happen in schools. It is at this point that the evaluation identifies key components of the selected model along with an expected timeline for the process to take hold in the school. Research at schools that have put comprehensive efforts into place has shown that one of the major roadblocks to the success of any program is getting the program widely and consistently used by staff around the school.

Before the evaluator can begin to collect information about implementation, the school will benefit from some common understandings about the stages staff typically go through to implement a new program. This discussion is likely to be most productive when grounded in a research-based theoretical framework—that is, when the staff has a



From *Taking Charge of Change* by Shirley M. Hord, William Rutherford, Leslie Huling-Austin, and Gene E. Hall, 1987

Figure 1. Levels of use related to instructional implementation

common vocabulary based upon research-proven concepts. Such a framework can promote meaningful dialogue about evaluation. Using a framework increases communication about both evaluation and implementation. One framework that works well is Levels of Use, developed by Shirley Hord and her colleagues (Hord, Rutherford, Huling-Austin, & Hall, 1988). A brief description of the levels that staff members go through as they work with an innovative program is explained in the box below.[4]

The levels of use Hord describes are as follows:

Non-Use: Teacher has little or no knowledge of the new approach, no involvement with it, and is doing nothing toward becoming involved.

Orientation: Teacher is acquiring information about the new approach and/or has explored its value and its orientation, what it will require.

Preparation: Teacher is preparing for first use of the innovation.

Mechanical use: Teacher starts to use the new approach but focuses her or his effort on the short-term, day-to-day use of the innovation with little time for reflection; use is disjointed and superficial.

Routine: Teacher use is stabilized. Few if any changes are being made in ongoing use. Teacher no longer needs to prepare or give additional thought to use this approach. Time is not spent improving the approach or identifying its consequences.

Refinement: Teacher varies the approach to increase impact. Teacher examines both short- and long-term consequences to learn more about what works best. Use of this approach is based on input from (and in coordination with) colleagues. It is at this point that the primary focus becomes benefiting students.

Integration: Teacher uses approach with related activities to achieve a collective impact on students. Teacher explores major modifications of the approach to ensure maximum benefit.

Renewal: User moves toward a new approach.

Additional information on how schoolwide information about Levels of Use can be summarized is available in Implementation Evaluation Transparency #5 and Implementation Evaluation Handout #2.

Developing Evaluation Questions

Once the preparatory work is done, schools should consider what kinds of information would help ensure complete program implementation. To do that, schools need to learn more about existing conditions. They should collect baseline information that paints a clear picture of the pace and scope of change taking place in the school.

[4] These concepts can be introduced to the staff using Implementation Evaluation Transparency #3, Levels of Use Related to Instructional Implementation. Implementation Evaluation Transparency #4 illustrates how the level of staff use can be assessed through a series of simple questions.

Review Existing Data

Those working on evaluation should start with a review of descriptive information about the school. This would include brief descriptions of program participants, an overview of the plan and goals for the comprehensive program, and contextual information. Much of this information can usually be pulled from a grant application, but it may need updating and further specification.

Decide What Additional Data To Collect

At this point, schools will begin developing research questions. The questions are written for two purposes: first, to explore concerns or issues, and second, to confirm hypotheses or troubleshoot problems. There is no set of generic questions that will work for all programs. Unique questions need to be written for each program to focus the data collection on:

- The type of program being implemented
- What the school is trying to accomplish
- Specific contextual issues facing the school

Sample implementation evaluation questions are shown in the box at the top of this page.[5]

Are staff members knowledgeable about comprehensive changes required by the reform model being implemented?

Do staff members demonstrate a commitment to the needed training?

Is the program being implemented as it was designed?

Are staff using the new instructional practices that were taught to them during inservice sessions?

While there are no magical questions that will work in all situations, there are criteria that can be applied to determine if the questions chosen will be useful in guiding the evaluation design. Questions should be:

- Clear
- Specific
- Pertinent to essential aspects or components of implementation of this program
- Focused on a manageable set of issues

The wording of these questions is a very important part of the process of designing an evaluation. How these questions are stated will have implications for the kinds of data that will be collected, the sources of the data, and the analyses that will be done on the data. Ultimately, the way the questions are worded will affect the kinds of conclusions that can be drawn about the program.

Here is an example. Suppose your implementation evaluation is geared to find out how clear the new program is to teachers. This issue of clarity can be addressed in several different evaluation questions. Here are two possible question formulations:

- Do the participating teachers have a clear understanding of the purpose and goals of the program?
- Have criteria been established to determine if the program is clear enough to the teachers so that they can implement it?

The data collection approach differs dramatically depending on which of these questions is chosen. For question #1, the evaluator would collect data from the teachers themselves to determine their understanding. But with question #2, the evaluator would be more likely to turn first to the program developer and to written documentation to learn if criteria existed and then, using this information, would design an instrument to be used with the teaching staff.

How questions are generated is very important. Schools should carefully consider who should be involved and what resources they should use. Without a doubt, the best evaluation work is done when multiple perspectives are taken into account. While staff may formulate a set of initial questions, many other stakeholders should have an opportunity to provide input. This will increase ownership and participation in the evaluation and increase the likelihood that evaluation results are used.

[5] These are displayed on Implementation Evaluation Transparency #6.

So how are these evaluation questions developed? One strategy is to interview program staff and then use their input to propose several evaluation questions for staff review. Another approach is to hold a meeting with staff to talk about the work that will be done throughout the school and then ask the staff to list their concerns about the program. This information can then be shaped into evaluation questions. Evaluation questions can also flow from an understanding of the factors that are most likely to help and hinder implementation. These are summarized in Implementation Evaluation #7. Remember that evidence for the evaluation can take many forms, but that the data collected must be relevant to program improvement decisions. Evaluators should ask themselves, "If staff knew the answer to this specific question, how could or would they act with this information?" Certain types of information, while interesting, may not help the staff to make changes. So the useful rule of thumb is to determine which data are most needed to correct or fine-tune a program.

Evaluators can also collect data related to factors that may be preventing program implementation, along with some documentation of ways that these barriers are being addressed. A simple form for this type of documentation is shown in Implementation Evaluation Handout #3. Program mangers are encouraged to plan intermittent review of such barriers to learn if adequate support is being provided for program implementation.

Because comprehensive reform is complex, it is important not to narrow down the data collection too early. Also, it is best to save all data collected. While some data may not seem immediately relevant, new issues may emerge during the course of the analysis phase, or program priorities may change.

Planning the Evaluation

Step 1: Work closely with the planning team and with the professional developers who are presenting training related to the school's comprehensive model. Knowing what staff will be learning and when they will be learning it is a crucial part of the implementation evaluation. In addition to the actual staff development days, there may be follow-up meetings and/or a series of benchmarks that establish the timeline for implementation. Staff need to be intimately familiar with this schedule and to use this information in evaluation design and measurement selection.

From the beginning, the evaluation must be structured around:

- The schedule of training events
- Key information that will be provided at each professional development event or meeting
- Likely stages of implementation (including information about typical variability among the staff in the pace of implementation)

Once this information has been gathered, it is time to sketch out the data collection design.

Professional developers are often excellent sources of detailed information about implementation of their model. They can provide information about other schools' experiences with the model and about problems that may crop up. They can describe program idiosyncrasies, such as whether teachers in certain grade levels are most likely to implement the program; whether certain trainings need repetition and support before teachers will adopt the approach; or what level of staff preparedness and support is needed for full implementation. A conversation with the professional development team can provide solid background for the evaluation plan.

Step 2: Design a matrix that lists the kinds of data that would answer the research questions and that pinpoints the best time to collect each kind of data. There are several things to consider in the design of the matrix: (1) how to ensure that you have adequate information, (2) how data collection will be conducted, and (3) when and where the data collection activities will occur. One of your goals will be to gather data from enough sources to provide balanced information.

This is the time to consider a variety of data-gathering strategies. When you are deciding which data to collect and how, there will be pragmatic considerations:

- The value these data have as evidence
- The cost to collect them
- The amount of intrusion into school routines
- Any ethical considerations or constraints being placed on the evaluation

Throughout these early stages of implementation evaluation, evaluators should keep the following key points in mind:

1. Encourage continuous reflection and thinking about the reform process.

2. Recognize there is no one-size-fits-all comprehensive reform model. Help staff realize that any reform model needs to be adapted for use at each school, and that input from staff is imperative. To ensure that progress is made, evaluation planning needs to include a timeline of events or activities as well as a description of what teachers are expected to implement during the year.

3. Inform the staff that for a school reform effort to be comprehensive, it needs full participation from a broad base of school community members. Including a greater number of stakeholders in evaluation planning encourages greater participation in the reform.

There are other considerations as well. One is how to communicate information about the evaluation to all participants. Duration of data collection, as well as coding and storage of data, are other concerns that will affect staff and program design.

A number of examples of data collection procedures are shown on Implementation Evaluation Transparencies 8, 9, and 10. Reviewing these ideas will provide some examples of procedures that are often used in implementation evaluation. Obviously the design developed for each school will need to consider the size of the school, the amount of time that staff have available for interviews, the structure of faculty meetings, and the timeline established for professional development.

Implementation Evaluation Transparency #11 illustrates the development of a matrix demonstrating data collection procedures for an elementary school. Once a matrix has been completed for your school, the matrix will serve as a visual representation of the evaluation design. It can serve as both road map (to show where the evaluation is headed) and timeline (to keep the data collection on schedule).

Step 3: Select tools that will provide you with answers to your evaluation questions. Be sure to consider a variety of data collection tools. In the selection of data collection tools, staff gathering data should keep several considerations in mind:

- Balance
- Validity and reliability
- Participant perceptions

It is often cost-effective to use preexisting instruments. These should be reviewed to make sure they are relevant to the school's needs.

To ensure unambiguous interpretation of data, it is important to pretest the items—that is, try them out with a number of staff members. Questions should elicit complete answers that directly address your questions.

To ensure practicality of design, schedule time not only for the data collection but also for the analyses and reporting of data. A general rule of thumb is that it takes one and a half to two times as much time to analyze the data as it does to collect them. It is also important to choose approaches that are simple enough to complete within the time available. If the evaluation has four days of data collection time available, for example, it will be impossible to schedule three days of interviews along with two days of focus-group meetings.

Collecting Evaluation Data

Data collection can include information about many components of a comprehensive program such as:

- Professional development activities
- Parental involvement
- External technical support and assistance

When collecting data, staff members need to accurately record what they see and hear and avoid making judgments. They should concentrate on recording observations or conversations in an objective way. To capture the information as cleanly as possible, the evaluation should include the development of data collection guides—forms providing questions and space for recording verbatim notes from interviews or classroom activities.

Data collectors should encourage reflective thinking by:

- Using wait time
- Keeping good eye contact
- Asking staff to explain their comments or to provide specific examples or anecdotes

Minimizing Bias

Bias is always an issue in data collection. To avoid getting a biased view of the program, data collectors need to ensure they are getting a broad representation of views. Therefore, it is essential to randomly select individuals to interview or observe, but at the same time to make sure that all key groups are included in your sample. Here is how this works. Suppose that one key aspect (variable) under study is how well teachers at grade level are implementing the new program. To avoid disruption, the evaluation could just ask for teachers to volunteer to participate in an interview, but then researchers would only get the staff members who already had a reason to share their perspective. To avoid bias in this case, while at the same time ensuring that each grade level is represented, the evaluation design should call for the random selection of one teacher from each grade level.

Bias is the personal and unreasoned distortion of judgment. Bias is evident when conclusions are reached, not based upon facts, but instead because those analyzing the data already have certain viewpoints or perspectives.

Those participating in the program are most likely to view the results as biased when an evaluation is unduly influenced by, disrupts, or threatens ongoing social and institutional relationships. If informants have a reason to distrust the evaluation process, they may appear helpful but can be withholding or shaping information out of self-interest.

To reduce the effects of bias during data collection:

1. Use unobtrusive measures whenever possible.

2. Make sure the purpose is completely clear to informants. Make certain they have a copy of your research questions, remind them why the evaluation is being done, and tell them what you will do with the information. This builds trust.

Reviewing the data and generating hypotheses about what they say may be the job of a small group. But getting a complete understanding of the underlying meaning often becomes a whole-group task. Structuring meeting time to encourage group input provides multiple perspectives while at the same time providing immediate feedback to a large number of stakeholders.

In particular, be sure to include those who spend their days implementing the program in any data interpretation activities. When staff members work with the data, they become familiar and comfortable with them. Making the findings more accessible to the staff increases the likelihood the results will be used.

3. Include dissidents and "cranks" to achieve a balanced picture.

4. Triangulate (checking your research question[s] against other already validated measures) with several collection methods.

5. If you sense you are being misled, focus on why.

6. Show field notes to an outside reader (without breaking confidentiality).

7. Keep your research questions firmly in mind.

Analyzing and Interpreting the Data

Once the data have been collected, the school staff must make sense of them. Meaning will emerge from analysis that is both systematic and thoughtful. The analysis requires blending technical skills to organize data quantitatively with intuitive skills to tease out the messages that may lie hidden behind the responses of individuals.

While the choice of analysis method depends on the type of information and the purpose of the analysis, the summary that emerges needs to describe either quantitative (percentages, averages) or qualitative (description of themes that emerge from the reader's point of view) information.

It's best to organize the data for each research question separately. This ensures that data addressing one question can be examined without contamination from data addressing other questions. By looking at all the data related to one question, data analyzers can determine if the data support one conclusion, or if in fact there are various perspectives. When reviewing the data, the evaluation should look at the "big picture," as well as smaller themes that surface. Once the data have been examined for each question, the analysis should expand its focus to include the full data collection. The staff may then begin to see a pattern of issues that touch multiple aspects of the program.

The next question to ask is what these data say about the path the program should take. Knowing what decisions are to be made and by whom will help determine the best way to conduct a secondary analysis of the data. If the staff wants to know how much time teachers at various grade levels need to cover certain material, then collecting unit completion information from every classroom teacher would prove most useful. If the principal wants to know whether teachers are adapting program components to provide different instruction for separate groups inside a classroom then the within-classroom variability data should be disaggregated to isolate findings for those subgroups.

Reporting the Data

It is likely that funding agencies will require some type of written report in a format that is useful to them, but addressing your report to their requirements alone leaves important work undone. Sending that evaluation report off to the funder's files will not improve your program. Instead, the findings of the evaluation need to be portrayed thoughtfully in a way that will communicate with the staff of this school. In the case of implementation evaluation, a summary of the results along with some help interpreting the data is of utmost importance. This is best done in a combination of oral and written reports. Since those implementing the program are busy people, it is important to keep both types of reports short, allowing time instead for the users to discuss the reports' implications for their day-to-day work.

Reporting to school staff should reflect the concerns of the audience. What are they worried about? What information do they need to tackle their most pressing concerns? The information should be presented in language the audience can relate to and understand.

Evaluation findings should be shared at both school and community meetings. To make the presentation of the data more accessible and interesting, the presenter should:

- Get the audience involved by giving them a brief warm-up activity.
- Try to talk with, not at, the audience.
- Use conversational language and avoid technical words.
- Present the data in creative formats that will engage the audience. Use graphs and charts to make the presentation of information as visual as possible.
- Punctuate the presentation with audience questions that will encourage the program implementers to reflect on the data.
- Place nothing between presenter and audience. Don't stand behind a lectern. If possible, mingle with the audience.
- Use the names of the participants whenever possible, and encourage them to interact with one another.
- Smile and look relaxed.
- Use humor whenever possible.
- Use personal anecdotes and stories. These give the audience something to relate to and bring the presentation down to earth.

Timing is vital in the preparation of both verbal and written reports. Those generating a report need to know the program schedule. For example, when will the planners hold their meetings? When will staff development take place? When is staff likely to make program adaptation? Reports should provide enough detail to enable the staff to make midcourse corrections. The evaluation reporting cannot wait until the end of a year or the completion of the project.

The number and type of groups that will receive the information are also crucial considerations. Ideally, findings should be shared with anyone who participates in this program. Whenever possible, information should go to staff, students, parents, and the community. Sometimes it is useful to share information in several formats for the different audiences.

There are a number of ways to present findings, the most common being a written report. Such reports can vary greatly in style, depending on the audience. Style options include journalistic summary, dialogue, testimony, question and answer, or scenario. Certain kinds of data may best be presented in a graph or chart, case studies, panel discussions, or simulations.

Presentation method and style should be tailored to the audience and their intentions—that is, who will receive the report and how they will use it. While the formal report may take longer, a draft of several key findings could be completed and distributed very quickly. For some audiences, small segments of findings doled out a bit at a time or a streamlined version of overall findings may suffice. But those who are working to implement the comprehensive program will benefit most from a report that is rich in descriptive detail.

School staffs are most likely to use the findings if:

- They have been closely associated with the evaluation effort
- They have a long-standing commitment to the use of data
- Conclusions are presented in a straightforward, understandable way
- They receive the information at the time they need it
- Evaluators share their ideas in draft form, solicit feedback, and make revisions

If possible, the written report should include comments and quotes from staff and/or students to make it more engaging to teachers. Staff or student comments lend credibility to the findings and give the information a human dimension.

Using Data To Make Program Improvements

To ensure that the data will be used, the evaluation effort also needs to include ways to facilitate discussion with decisionmakers about the steps they will take to put the data into action. These ideas should be included in a school improvement plan that lays out strategies to strengthen instructional practices. The plan should be clear about what teachers are expected to do, include activities that are an integral part of daily instruction for all teachers, and ensure that teachers have or develop the skills to implement changes.

Once the leadership team or a steering committee has the data in hand, take these steps:

- Review the strategies and action steps originally proposed in your grant application or school improvement plan. Identify who was responsible for implementation and ascertain how far along the school was supposed to be at the time of data collection.
- Use the data to identify the parts of the plan or programs that are not being implemented and other challenges facing staff.
- Make sure staff are aware of the findings and then ask what else could be done to help your school make changes.
- With staff input, determine what additional training is needed to improve the implementation process. Decide what kind of staff development can get this done.
- Determine if new materials are needed and how they will be purchased or developed.
- Determine how to provide ongoing support to sustain implementation of the plan.
- Determine what added resources are needed to implement the revised improvement plan and how they will be obtained.

- Reestablish responsibilities and timelines needed to implement the revised plan.
- Communicate what has been incorporated into the plan to all staff, and ask all staff to take action.
- Review the implementation evaluation design. Make changes as needed to gather data reflecting the modifications.

Reflections About the Implementation Evaluation Work

The process of doing implementation evaluation may often seem paradoxical for those involved in program implementation. In most schools, a mix of insiders and outsiders is likely to be involved in the evaluation. While this can strengthen the process, it also adds to the complexity. Often, the insiders work closely with those implementing the comprehensive program, while outsiders bring the perspective of impartial observers of change. The combination of these two perspectives adds richness to the process but also requires openness and sensitivity about the working relationships between the groups.

In addition, members of the evaluation team need to take on different roles at various times. Data collectors are asked to be equally comfortable talking with those in authority and those who have very little formal power. They must recognize that comprehensive programs need the input of both groups if they are to succeed. When determining if implementation is occurring, data collectors need to be genuinely invisible, quietly watching. But when the time comes to communicate results, the same individuals need to be highly visible, sharing important information and explaining the findings.

For all these reasons, selecting who to serve on your evaluation team is an important decision. And determining who should report the results is also a critical decision. Presenting results can at times bring out tensions between two opposed groups: those who are working to get a new program in place and those resisting or struggling. The sensitive presentation of formative evaluation data has the potential to open the lines of communication between these two.

Summary

Implementation evaluation is a way to assess the work between program planning and program impact. Planning and conducting a formative evaluation is an ambitious project since it requires data collection that reaches into the classroom. To target the evaluation, it is necessary to develop specific evaluation questions, identify the most appropriate sources for the data needed, organize to get broad participation in the data collection and analysis process, and determine the best time and place to summarize the data and report the findings. Conducting an effective implementation evaluation means keeping in close contact with the implementation process and the staff members who are making this program a reality.

Resources

Beyer, B.K. (1995). *How to conduct a formative evaluation.* Alexandria, VA: Association for Supervision and Curriculum Development.

This book describes how to conduct a formative evaluation of educational programs by assessing the program during various stages of its development. The author provides practical checklists, data-collection instruments, and other resources to assist in conducting the evaluation.

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

This comprehensive guide offers educators step-by-step procedures and practical guidance needed to conduct sensible assessments and evaluations, and record and measure progress. It also instructs the reader on how to use evaluation information to aid in school planning and improve management decisions.

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

This book is one component of the Sage Publications series, *The Program Evaluation Kit,* a set of guidebooks written to guide and assist program evaluators in planning and managing evaluations. The guide will help practitioners plan an evaluation of program implementation and design, and use appropriate instruments for generating data to support the plan. Procedures in the "how to" sections of the book are presented step by step to give maximum practical advice.

References

Bodilly, S.J., Keltner, B., Purnell, S.W., Reichardt, R.E, & Schuyler, G.L. (1998). *Lessons from New American Schools' scale-up phase: Prospects for bringing designs to multiple schools.* Santa Monica, CA: Rand.

Buechler, M. (1997). *Scaling up: The role of national networks in spreading education reform.* Unpublished manuscript. Portland, OR: Northwest Regional Educational Laboratory.

Hord, S.M., Rutherford, W.L., Huling-Austin, L., & Hall, G.E. (1987). *Taking charge of change.* Alexandria, VA: Association for Supervision and Curriculum Development.

Instructions for Implementation Evaluation Transparencies

The transparencies in the Overview section provided background information on the issue of formative or implementation evaluation, including an outline of the purpose of this type of evaluation (Overview Evaluation #2), comparisons of formative and summative evaluation (Overview Evaluation #4), and generic formative evaluation questions (Overview Evaluation #7). Each Implementation Evaluation transparency discusses issues that arise early in the evaluation process as formative evaluation design is being generated.

Transparency #1

Outlines several areas of misunderstanding that staff can have about the evaluation process. Because these can undermine data collection during formative evaluation, the evaluator might use this transparency to initiate a brief discussion with staff to clarify any misconceptions.

Transparency #2

Lists advice to those who will be planning and conducting implementation evaluation for a comprehensive program.

Transparency #3

Summarizes the Levels of Use framework, which shows that staff move through a number of phases before they can effectively use a new approach. However, their progression through these levels of use is not uniform, and without support many of the staff will not make it all the way around the circle. Many staff struggle with the mechanical use of a new program. Then, because they lack additional support and encouragement, they drift into routine use of the approach. Explain to the group that it is really in the refinement phase of implementation that student benefits are noted. (The handout on the Levels of Use provides a brief description of each of the various levels.)

Transparency #4

Explains that interview questions such as the ones shown on this transparency can be used to determine where a staff is in relation to its use of a new approach.

Transparency #5

This transparency is meant to accompany the handout on program components, so that the presenter can explain the structure of this type of data summary. A unique matrix for each comprehensive program is developed by working closely with the program staff to identify the key components to be implemented. Following the development of the list of essential program components, data on how completely each component is being implemented by each teacher in the school are gathered via interviews and observations. When all data are collected, the pattern of implementation for the whole school is displayed in the matrix as illustrated in the handout. (This handout displays findings for 10 teachers in the building.) When showing this transparency, the presenter needs to explain that this transparency only shows the findings for the first component. In this row, each one of the asterisks represents the current level of implementation of one teacher in the building. This particular pattern shows that one of the 10 teachers has not yet rearranged the classroom (the first essential component of the program), and one of the teachers has progressed to the point of refining the process of classroom rearrangement to maximize effectiveness. The implementation level of the remaining eight teachers is somewhere in between.

Transparency #6

Provides some sample evaluation questions—ones that might be developed early in the process of comprehensive reform.

Transparency #7

This transparency outlines a number of factors that have been shown to affect program implementation.

Transparencies #8 and #9

These two transparencies list a number of sources of data for the implementation evaluation.

Transparency #10

Provides additional explanation of the types of questionnaires or interview data that could be collected.

Transparency #11

Illustrates a data collection matrix displaying evaluation questions, data sources and timelines, and approaches for collecting the needed information.

Instructions for Implementation Evaluation Handouts

Handout #1: Levels of Use About Instructional Implementation

This handout provides a concise list of the Levels of Use, which characterizes the implementation and innovation. Staff start at level 0, where they have no knowledge of the changes they are being asked to make in a comprehensive reform model, and then proceed through the orientation and preparation levels. When staff first begin to use a new instructional approach in the classroom, they are entering the mechanical-use stage where they need both feedback and support. If these are not provided, staff may continue to use the new instructional approach but will slip into routine use. When using the innovation in a routine way, staff are less likely to get the full benefit of the new approach. Ideally, staff need to be helped to move to the refinement level, where they make adjustments that provide the greatest benefit to the students.

Handout #2: Program Components

This handout illustrates how the Levels of Use can be used in program evaluation. The evaluator needs to identify the key components of the comprehensive reform model that are to be implemented at this site and list these on a form like this. Then, by interviewing the teachers at the school (using questions like the ones displayed on Implementation Evaluation Transparency #4), the evaluator can assess how far along the various staff members are in putting these new practices into place. The information gathered can be displayed on a grid like the one in this handout, without violating confidentiality. For example, this handout demonstrates that all 10 teachers in this school are mechanically preparing their units collaboratively. However, when it comes to another program component (instruction is resequenced to match assessment expectations), two of the 10 staff (20 percent) are at the refinement phase (making adjustment in the classroom) and the remaining 80 percent of the teachers are struggling at the mechanical use level, with 60 percent just beginning mechanical use and another 20 percent reaching more advanced levels in their application of this approach. The purpose of a chart like this is to demonstrate progress toward implementation and to illustrate specific areas where additional support or staff development are needed. For example, the data on this handout demonstrate that teachers probably do not need added training on rearranging the classroom.

Handout # 3: Documentation of Implementation Interference

This handout illustrates how an evaluator can record various events in the school that have an impact on implementation. The first column of this matrix lists a number of general areas where issues can arise that interfere with comprehensive reform. Evaluators are likely to learn of these issues during interviews with staff or visits to the school. In the second and third columns, the evaluator would list the specific problem that was noted in the general area and the source of that information, along with the date that the concern was noted. This matrix can be shared with the program staff periodically as a way to determine if the interfering factors are being addressed. Program staff can be asked to indicate if they are aware of these issues and, if so, how the concerns are being addressed. New data should be gathered in the same way as the old data to determine if barriers are coming down. All this information can be recorded succinctly on a chart like the one in the handout.

Small-Group Activities

Each small-group activity is designed to reinforce or stimulate the discussion on a particular topic or concept. They may be conducted before or after the discussion. If the activity is done before the discussion, the topic should be briefly introduced first. As a presenter, you should guide the participants through the activity and then lead an interactive discussion of the results of the groups' work, drawing from the contents of the guidebook as appropriate to reinforce and/or enrich the discussion.

The small-group activity can also be scheduled to follow a more detailed discussion of the topic. In this case, the activity provides a way for the participants to apply what they have learned in the presentation and discussion.

Divide the audience into groups of about five people. The group can consist of members of a school team or just participants selected by various means to form a group.

As the evaluator introduces evaluation concepts to the staff, he or she can also begin to collect data about staff attitudes and beliefs. Four ideas for doing this are in this section. These activities can be used to spark staff conversations about the reform model and to help the evaluation planners understand the context for program implementation. The questions can be adapted to each site and used to collect formative evaluation evidence from staff. Once adapted, such questions can be used during interviews or during staff meetings. Following the activity, refer the participants to parts of the guidebook that discuss evaluation models and data collection (for instance, the Data Collection Matrix on Page 67).

Small-Group Activity #1

Staff input is also helpful in identifying site-specific issues related to the comprehensive nature of a program. To gather data about a program, the evaluator can encourage staff to discuss the benefits and limitations of the new program from their own perspectives. Staff meeting time can be used to get people to talk about the model they are adopting:

- What is the strongest feature of the model that you have chosen? What makes it strong? How will you know that it is having the desired impact?
- What is the weakest feature? How can you strengthen its impact?

Small-Group Activity #2

After staff have participated in professional development in which key components of the new model are revealed to the staff, the evaluator can conduct staff interviews to answer such questions as:

- How does this work connect with other work underway in the school? How much do programs overlap? How much will this overlap affect implementation? Is staff trying to implement several programs simultaneously?
- How much innovation and change does this reform demand of staff?
- How much does the project depend on help and support from outside the school?
  - From trainers?
  - From community members?
  - From students? (Attendance or willingness to put in extra effort?)
  - From outside funding? (This is related to project sustainability.)

Small-Group Activity #3

When conducting an evaluation for a comprehensive program, the evaluator needs to determine if some aspects of the system limit progress. To identify what might slow down program implementation, the evaluators can start zeroing in on this information early in the process. To help secure information about systemic issues, the evaluator can ask staff about systemic barriers that prevent program implementation.

To do so, the evaluator might ask the staff to fill in the survey below:

What parts of the system (school, district, state, or community) might limit the school from using this new approach? List those limitations below, then rate the seriousness of these limitations on a 1 to 5 scale:

1 = least serious
5 = most serious

- School barriers
  1 2 3 4 5
- District barriers
  1 2 3 4 5
- State barriers
  1 2 3 4 5
- Community barriers
  1 2 3 4 5

Small-Group Activity #4

This activity is designed to be used as the model is being implemented. The evaluator can ask each individual in the group to complete his or her own personal rating on these items and then to work in small groups to reach consensus.

To introduce this activity, the evaluator can tell the school staff that reform models work best in situations that have open lines of communication. This enables consistent implementation of the key elements of any reform model. Because it is difficult for any model to get all staff to "buy in" to the project, it is helpful to get staff perspectives as the model is being implemented. This activity asks staff to help improve the work of the school by critiquing and rating the work in progress.

Ask all participants to rate (on a 1-5 scale) how well they believe the school is doing in certain areas such as the following:

1 = doing poorly
5 = doing well

- Being clear about what the end result of the program will be for students
  1 2 3 4 5
- Promoting teamwork and opportunities for staff to learn from one another
  1 2 3 4 5
- Having a shared vision about how the new program will operate
  1 2 3 4 5
- Knowing the role of each staff member in the project
  1 2 3 4 5
- Having all staff use the same instructional practices
  1 2 3 4 5

Once everyone has done the ratings individually, take 10 minutes of the staff meeting time to form small groups, asking staff members to compare their ratings and to discuss how they will know that they have achieved these various expectations for the project.

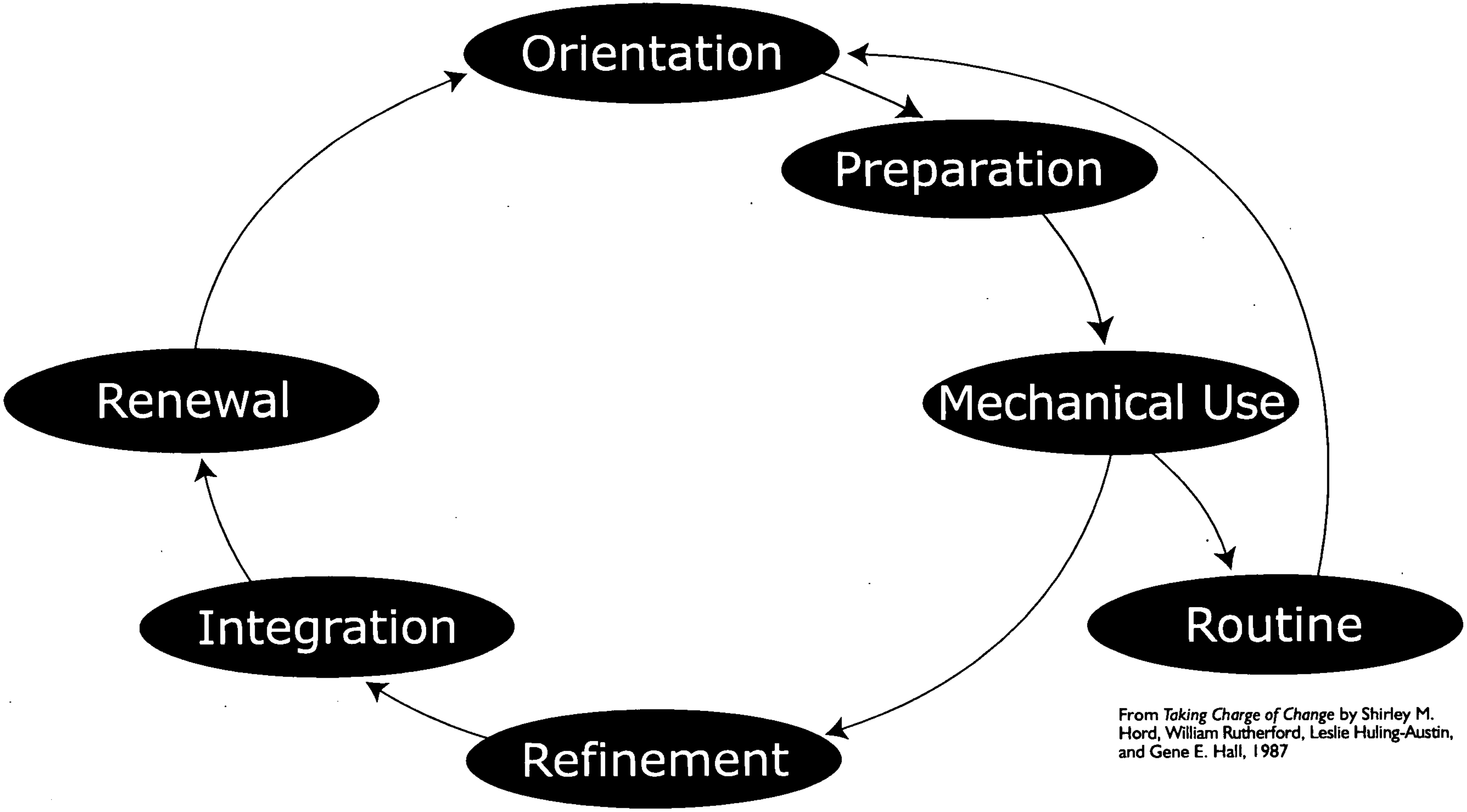
Possible Misconceptions About Program Evaluation

- Program (not staff) is being evaluated
- Data will be used to improve program (not to remove funding)
- Evaluator needs to be connected to (not isolated from) the program staff
- Evaluation is a helping profession

Data Collection Considerations in Formative/Implementation Evaluation

- Get involved early
- Incorporate ideas from program staff
- Take total school context into account
- Learn how program "plays out" in the classroom
- Provide feedback ASAP

Levels of Use Related to Instructional Implementation



Orientation

Preparation

Mechanical Use

Routine

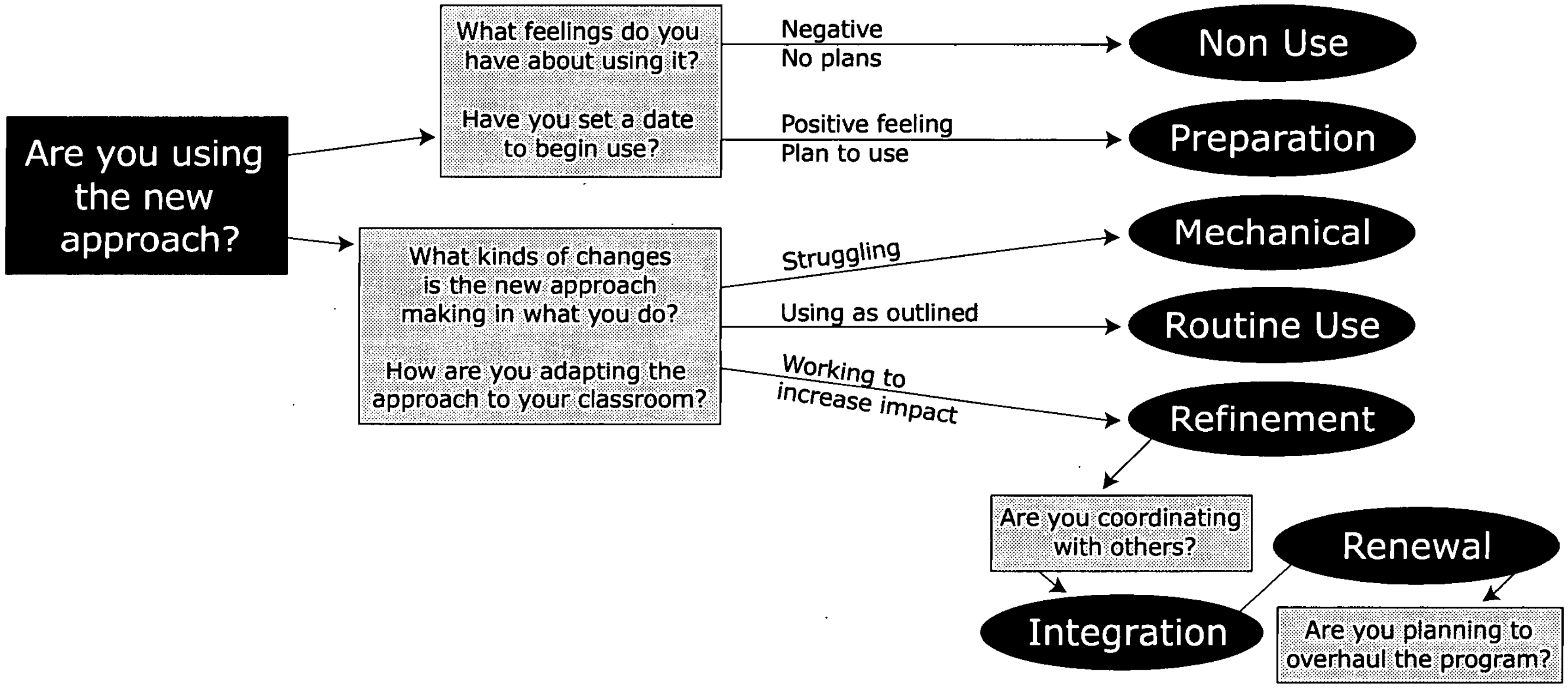
Refinement

Integration

Renewal

From *Taking Charge of Change* by Shirley M. Hord, William Rutherford, Leslie Huling-Austin, and Gene E. Hall, 1987

Interview Questions To Assess Implementation



Are you using the new approach?

What feelings do you have about using it?

Have you set a date to begin use?

What kinds of changes is the new approach making in what you do?

How are you adapting the approach to your classroom?

Negative
No plans

Positive feeling
Plan to use

Struggling

Using as outlined

Working to
increase impact

Non Use

Preparation

Mechanical

Routine Use

Refinement

Are you coordinating with others?

Integration

Renewal

Are you planning to overhaul the program?

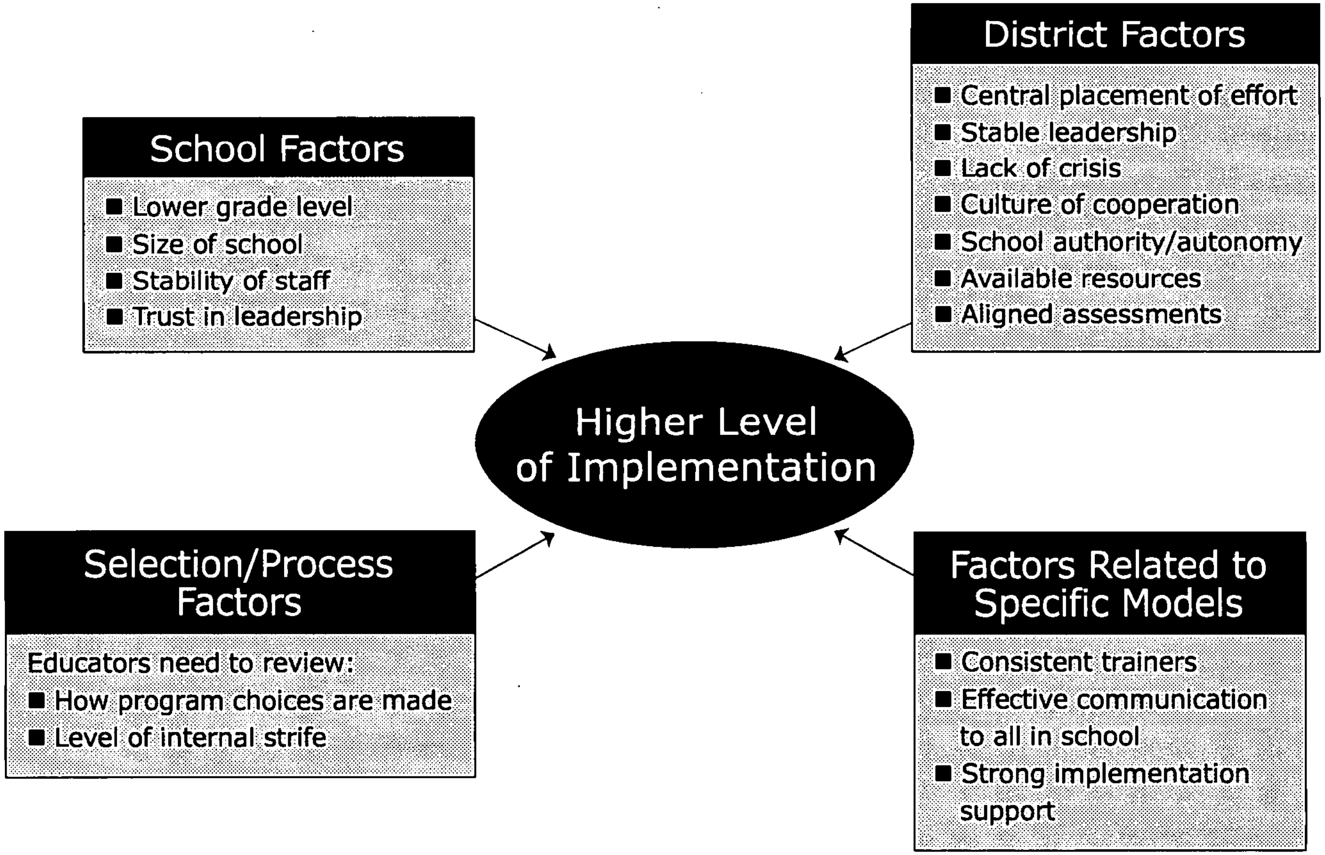
Program Implementation for Teachers in a Building

	NON USE ↓		PREPARATION ↓	MECHANICAL USE ↓	ROUTINE USE ↓	REFINEMENT ↓
PROGRAM COMPONENTS	0	1	2	3	4	5
1. Classroom has been rearranged to facilitate implementation	10% *		10% *	40% ****	30% ***	10% *

Sample Formative Evaluation Questions

- Is staff knowledgeable about comprehensive changes required by the reform models being implemented?
- Does staff demonstrate a commitment to the needed training?
- Is the program being implemented as it was designed?
- Is staff using the new instructional practices that were taught to them during inservices?

What Helps Implementation



Higher Level of Implementation

School Factors

- Lower grade level
- Size of school
- Stability of staff
- Trust in leadership

District Factors

- Central placement of effort
- Stable leadership
- Lack of crisis
- Culture of cooperation
- School authority/autonomy
- Available resources
- Aligned assessments

Selection/Process Factors

Educators need to review:

- How program choices are made
- Level of internal strife

Factors Related to Specific Models

- Consistent trainers
- Effective communication to all in school
- Strong implementation support

Exploration of Tools for Evaluation

Documents or artifacts from teachers—Records from classroom teachers: gradebooks, lesson plans, attendance, etc.

Documents or artifacts from students—Homework, assignments, quizzes, projects, etc.

Observation—Watching for specified classroom behaviors, transactions, or arrangement.

Questionnaire survey—Polling to assess opinion, attitude, or ideas among students, teachers, or community.

Examples of Data Collection Techniques

Artifacts

- Collect copies of student assessments or homework being used by the teachers
- Review minutes of staff meetings
- Review grade books, lesson plans

Observation Records

- Take notes about decisions made during cross-teacher collaborative meetings
- Describe classroom organization

Examples of Data Collection Techniques Related to Implementation

Questionnaire/Interviews

- Send out a questionnaire following a staff development activity
- Ask teachers to check off units that they have completed
- Talk with students to determine if the materials are being used in the classroom
- Ask teachers for specific examples of how they integrate their curriculum
- Ask teachers to describe one thing they did this week to put the model in place

Data Evaluation Matrix

Evaluation Questions	Data Source(s)	Data Collection Procedures	When and Where Data Will occur
Are staff members knowledgeable about comprehensive changes required by the reform model being implemented?	Staff	Interviews Staff surveys	Late October staff meeting and in individual teacher classrooms
Do staff members demonstrate a commitment to the needed training?	Principal Staff Developers	Interview with principal Records of staff attendance at optional inservice activities	November Following December training sessions
Is the program being implemented as it was designed?	Staff and students	Classroom observations using instruments developed with staff input Focus groups with students	Late March One hour focus group each with 4th, 5th, and 6th graders
Is staff using the new instructional practices that were taught to them during inservices?	Inservice training materials Staff	Staff interviews Gather and analyze a sample of lesson plans and class handouts	Once each month from December to May

Handout: Levels of Use About Instructional Implementation

0. Non Use

Teacher has little or no knowledge of the new approach, no involvement with it, and is doing nothing toward becoming involved.

I. Orientation

Teacher is acquiring information about the new approach and/or has explored its value and its orientation, what it will require.

II. Preparation

Teacher is preparing for first use of the innovation.

III. Mechanical use

Teacher starts to use the new approach, but focuses his or her effort on the short-term, day-to-day use of the innovation with little time for reflection; use is disjointed and superficial.

IV. Routine (a)

Teacher use is stabilized. Few if any changes are being made in ongoing use. Teacher no longer needs to prepare or give additional thought to use this approach. Time is not spent improving the approach or identifying its consequences.

IV. Refinement (b)

Teacher varies the approach to increase the impact. Teacher examines both short- and long-term consequences to learn more about what works best. Use of this approach is based on input from (and in coordination with) colleagues. It is at this point that the primary focus becomes benefiting students.

V. Integration

Teacher uses approach with related activities to achieve a collective impact on students. Teachers explore major modifications of the approach to ensure maximum benefit.

VI. Renewal

User moves toward a new approach.

From *Taking Charge of Change*, Shirley M. Hord, William Rutherford, Leslie Huling-Austin, and Gene E. Hall, Program Components

Handout: Program Components

	PREPARING TO USE		MECHANICAL ALL USE		REFINING DAY-TO-DAY USE	
PROGRAM COMPONENTS	0	1	2	3	4	5
1. Classroom arrangements have been made to facilitate implementation					90% ***** ****	10% *
2. Classroom environment assessed to determine who will facilitate implementation			20% **		20% **	60% *** ***
3. Teacher knowledge of students' interest guides program design			20% **	60% *** ***	10% *	10% *
4. Teachers prepare units in collaboration with others at their grade level			100% ***** *****			
5. Basic skills integrated into instruction			60% *** ***		40% ****	
6. Picture books are used as recommended	10% *		30% ***	40% ****	20% **	
7. Students assess their own learning	10% *		10% *	40% ****	30% ***	10% *
8. Instruction is resequenced to match with assessment expectations			20% **	60% *** ***	20% **	

Handout: Documentation of Implementation Interference

Issue	Specific Information *Data Source*	Date Noted	How was concern addressed?	Improvement noted?
Finances	■ Coordinators' time cut back because of limited funds *Examining budget records*			
Leadership	■ Principal does not act like he or she values the program; does not attend staff development, says little to staff *Meeting observation*			
Commitment	■ No pressure for commitment; teachers can choose to implement program at whatever level they wish *Teacher interviews*			
Political Issues	■ Administrators make decisions based upon political pressure *Interviews*			
Group Conflicts	■ Staff diversity causes internal conflicts			
Facilities	■ Building cannot be upgraded to allow technology needed for program implementation			
Management/ Communication/ Scheduling	■ Communication within the site is dysfunctional ■ Staff reschedule students throughout the year			

Impact Evaluation

This section of the guidebook addresses the question of whether the intervention (in this case, the implementation of a particular school reform model or approach) has made a difference at the school. For example, has it changed any school policy and practice? Strengthened instructional strategies? Improved student achievement? Has it contributed to the ultimate goal of providing opportunities for all students to meet high standards? The section presents several commonly used evaluation models, discusses advantages and disadvantages of each, and provides a step-by-step illustration of how each model can be implemented in a school setting.

The ultimate outcome we are looking for is improved student performance in academic subject areas, attitudes, and behavior.

It is common practice to use the terms "outcome" and "impact" interchangeably. In this section, we make a distinction between the two words. Outcomes will be used to refer to any results or consequences of an intervention—in this case, a whole-school reform effort. Impact is a particular type of outcomes. It refers to the ultimate results or outcomes. In the case of whole-school reform, we are really talking about results for students. For example, a whole-school reform effort can and usually does improve communication among the faculty and school administrators. It may also increase parental involvement with school activities. These are certainly desired outcomes. However, the ultimate outcome we are looking for is improved student performance in academic subject areas, attitudes, and behavior. These outcomes will therefore be considered as impact. For purposes of this section, we will use the term impact evaluation to include both outcomes and impact (Yap, 1997).

Outcomes can occur at many levels. We can assess outcomes at three interrelated levels: system, teacher, and student. At the system level, the intervention may have changed the way the school allocates resources and time for instruction. It may have affected its policy on professional development. At the teacher level, instructional strategies may have changed as a result of the intervention. Assessment practices may have been affected. At the student level, performance may have changed or improved on various measures.

Impact evaluation should be conducted only after a program has attained a sufficient level of stability. In practice, impact evaluation should be preceded by an implementation evaluation to make sure that the intended program elements have been put in place before we attempt to look at their effects. Assessing the impact of a nonentity—a program that has yet to be put in place—is meaningless and a waste of resources that can be put to better use (such as ensuring a high-fidelity implementation of the program).

Evaluation Models

The central question to be addressed in an impact evaluation is whether the intervention, in this case a whole-school reform effort, has made a difference for the target groups. There are of course different ways to find out whether the effort has made a difference. The different ways are sometimes described as evaluation models. The models can differ in many ways. An important difference is the extent to which the results they produce allow us to connect the implementation of various program elements with the outcomes or impact—to make a causal link between the two. This is sometimes described as the scientific rigor or validity of the model. In other words, some models are more likely than others to produce results that allow us to establish a causal link.

Some models are more likely than others to produce results that allow us to establish a causal link.

There can be as many models as there are program evaluators. However, the most commonly used models are: pretest-posttest model, comparison group model, regression model, and control group model. While the models are different, each must establish a standard or expectation

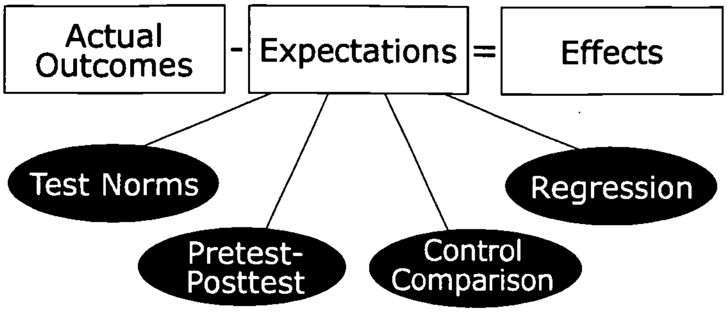


Figure 2. Evaluation model

against which to examine the program results. In other words, each must address this important question: What would be the expectation if the intervention was not implemented at the school? That is, how would students have performed without the program?

For example, in the pretest-posttest model, the expectation is that without the intervention, things will continue to go the way they have gone before. Teachers will continue to teach as they did before, and students will continue to perform as they did before. The baseline before the intervention will in fact be the expectation. Any difference, positive or negative, that occurs following the intervention is therefore attributable to the intervention (Tallmadge, 1982).

In the control/comparison group models, the standard or expectation is that without the intervention, things should be very much like those that exist in a similar or equivalent school or group of students. The critical issue is, of course, to identify and select an equivalent or similar school or group of students to be the control or comparison group.

The regression model uses a statistical method to predict or project what things would have been like without the intervention. The method takes into account most, if not all, relevant factors, including such things as current status and critical contextual variables (for example, demographic and socioeconomic backgrounds of schools and students).

For each model, once the no-intervention standard or expectation is set up, the actual state of affairs (instructional practice, say, or student performance) is then compared with the expectation. With varying degrees of confidence, we then attribute the difference to the intervention as illustrated above.

Each model, however, implicitly makes the "other things being equal" assumption. That is to say, other than the intervention—the whole-school reform effort—there is no significant difference between the project students and students used to set up the standard or expectation. This assumption, of course, is not always true. To the extent that this assumption does not hold, it is difficult to make a connection between program implementation and impact. In other words, it becomes problematic to attribute the outcomes or impact to the program.

Pretest-Posttest Model

This model makes the assumption that without the intervention, things will go on as they did before. Other things being equal, teachers will continue to teach as they did before, and students will continue to show the same pattern of achievement as they did before. With the intervention, things will change over time, it is hoped in a positive way.

This model assumes that the intervention occurs between pretest and posttest. Any difference that is detected between the two points in time will be attributed to the intervention. The model can include repeated measures. For example, both teaching practice and student achievement can be measured repeatedly at predetermined intervals (for example, twice a year or annually). The pattern of change at different points in time can then be interpreted as a result of the intervention. If the pattern of student achievement shows an upward trend over time (say, several years) then one can interpret the trend as evidence of sustained effects of the intervention (Blum, Yap, & Butler, 1991; Kushman & Yap, 1999).

Ideally, pretest and posttest measures should be taken from intact cohorts of students (the same students at two or more points in time). This is especially important when the in-

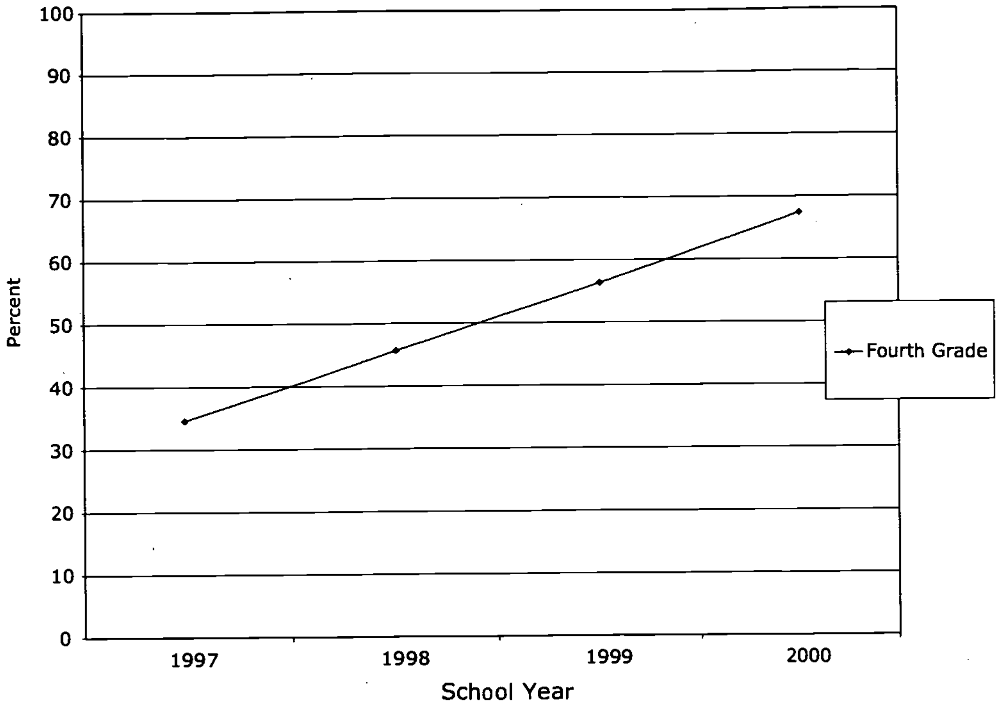


Figure 3. Pretest-Posttest model

tent is to measure gains of individual students. However, in a school setting, pretest and posttest measures, or repeated measures over a longer period of time, are typically taken from non-intact cohort groups. For example, assessments may be conducted with third-graders at a school on an annual basis. In this case, the measurements are obviously not taken from the same students. While the unknown bias that may result is a concern, it is less critical when we are primarily interested in knowing how a school, as a unit of change, is being affected by the intervention over time.

To get a longitudinal perspective, the pretest-posttest model can be implemented as a quasi-time-series model where repeated measures are taken over several years. For example, assessments can be conducted on an annual basis to identify longitudinal patterns and trends in student outcomes, as shown below. The line graph shows increasing percentages of students meeting state standards from 1997 (baseline year) through 2000.

Typically, program outcomes and impact are measured longitudinally over several years. A consistently positive or upward trend can provide compelling evidence that the intervention is producing positive results. It is, however, difficult to rule out completely the possibility that the positive trend is the result of some other factors (such as change in student population or change in teaching staff).

Implementation Steps. The pretest-posttest model is relatively easy to implement. Important steps include the following:

1. Decide what outcomes you want to look at

2. Select or develop instruments to collect the pertinent data

Advantages. The greatest advantage of the pretest-posttest model is that it is highly feasible in a school setting. It does not require a control or comparison group or a high level of statistical expertise to implement the model. It is one of the least intrusive models and it does not impose a heavy data burden on teachers and students. It can assess progress against a baseline. Further, it can measure growth or an absolute level of performance (Messick, 1985). For example, we can measure growth (an increase of 10 percent) toward meeting state standards. Alternatively, we can assess the extent to which an absolute level of performance (e.g., 60 percent of students meeting state standards for a particular school year) is attained.

Disadvantages. The greatest disadvantage of this model is that it lacks scientific rigor unless it is implemented as a true time-series model, using intact cohorts. In a true time-series model, the intervention is introduced and withdrawn at will or at random at various points in time. The assumption is that when the intervention is withdrawn at any point in time, things will revert to the preintervention status. In a school setting, however, it is seldom, if ever, possible to introduce and withdraw an intervention at will over time.

3. Decide whether sampling is desired

4. Administer the instruments to target groups at pretest time (for example, the beginning of school year)

5. Administer the instruments at posttest time (for example, the end of school year)

6. Analyze and interpret the evaluation data

7. Report findings to stakeholder groups

8. Use evaluation data for accountability and program improvement

The following example illustrates the use of the pretest-posttest model to assess the impact of a school reform model.

Pretest-Posttest Model—An Example

The Jefferson Elementary School has an enrollment of 500 students in kindergarten through grade five. The school has a very diverse student population with 35 percent minority students. Approximately 60 percent of the students are in the free or reduced-price lunch program. Jefferson has just adopted a comprehensive school reform model—Reading Enhancement—for schoolwide implementation. A school leadership team is formed to oversee the school improvement effort.

The statewide assessment program conducts testing of students in grades three and five in two core subject areas—reading and mathematics. The assessment takes place in April each year. The school also participates in districtwide writing assessment with grade five students in April each year.

The school leadership team wants to know if student performance is improving with the implementation of the school reform model. The team chooses to use the pretest-posttest model to conduct an impact evaluation of the school reform model. To take advantage of existing data available from the statewide assessment program, the team decides to use an annual testing cycle—April to April—rather than fall-to-spring to assess impact.

Step #1

The school leadership team, following extensive discussions with school staff, parents, and members from the community, decides to look at student performance in four areas: reading, mathematics, writing, and attendance. Even though the school reform model is focused on reading, the school and the community feel that it is important to look at other success indicators for the entire school.

Step #2

Most of the pertinent data will come from the statewide assessment program, including student achievement in reading and mathematics. Writing assessment data (for grade five only) will come from the district office. The only data collection instrument that needs to be created is a data form to provide summary data on student attendance—number of days absent per school year.

Step #3

Student achievement data are obtained—electronically when feasible—from the statewide assessment program for grades three and five. There are approximately 60 students in each of these grades. Data are obtained for all the students. No sampling is needed or desired. In addition, writing assessment data are obtained for all students in grade five. Attendance data are collected from school attendance records for all students in grades three and five. No sampling procedures are used.

Step #4

In the preceding school year, after receiving training in test administration from state-level staff as part of the statewide assessment process, the classroom teachers administer the criterion-referenced tests in reading and mathematics to students in grades three and five in April. The tests, which have been aligned with the state content standards, consist of multiple-choice items and a few open-ended items. The tests are scored by a vendor and the results provided to the school and district as well as the state department of education.

In addition, the writing assessment is conducted with students in grade five following procedures established by the district office.

Step #5

Also as part of the statewide assessment process, the criterion-referenced tests in reading and mathematics are administered to students in grades three and five in April in the current school year. In addition, the writing assessment is conducted with students in grade five following procedures established by the district office.

Step #6

A database is set up to store and manage all the data, including attendance data collected at the end of the school year. The database contains statewide assessment data (reading and mathematics) as well as districtwide writing assessment data for the current and preceding school years. The data are analyzed to provide percentages of students (grades three and five) who meet the state standards or benchmarks for the current school year and the preceding school year—prior to the implementation of the school reform model. A difference in percentage points provides an indication of impact. Attendance data are analyzed to provide an average (mean or median) number of days absent for each school year. Similar analyses will be conducted in future years to detect any consistent trends and patterns.

Step #7

Results of the analysis are provided in reader-friendly data displays (e.g., bar charts and line graphs) and easy-to-understand narratives. They are shared and discussed with stakeholder groups, including school staff, site council, parents, and members of the community.

Step #8

The results are provided to the district office and the state department of education to determine whether adequate progress has been made by the school. In addition, a meeting is held with the school leadership team, other key school staff, parents, and community members for an indepth review of the data to explore plausible reasons for the findings and to develop recommendations and an action plan for continuous improvement.

Comparison Group Model

This model provides an expectation of program outcomes based on a comparable group (Kushman & Yap, 1999). The comparison group, when selected appropriately, provides a basis for determining what might be expected to occur in the absence of the intervention. The comparison group should be similar (if not equivalent) to the intervention group in all relevant respects. Some of the pertinent factors include current achievement level, socioeconomic and related demographic factors, school locale, and size. Other things being equal, any detected difference between the two groups is attributable as impact of the intervention. The bar graph on Page 54 shows higher percentages of project students meeting state standards relative to their comparison counterparts in reading and mathematics.

Implementation Steps. Important steps in implementing the comparison group model include the following:

1. Decide what outcomes you want to look at

2. Select or develop instruments to collect the pertinent data

3. Identify and select a comparison group

4. Decide whether sampling is desired

5. Administer the instruments to both project and comparison groups

6. Analyze and interpret the evaluation data

7. Report findings to stakeholder groups

8. Use evaluation data for accountability and program improvement

The following example illustrates the use of the comparison group model to assess the impact of a school reform model.

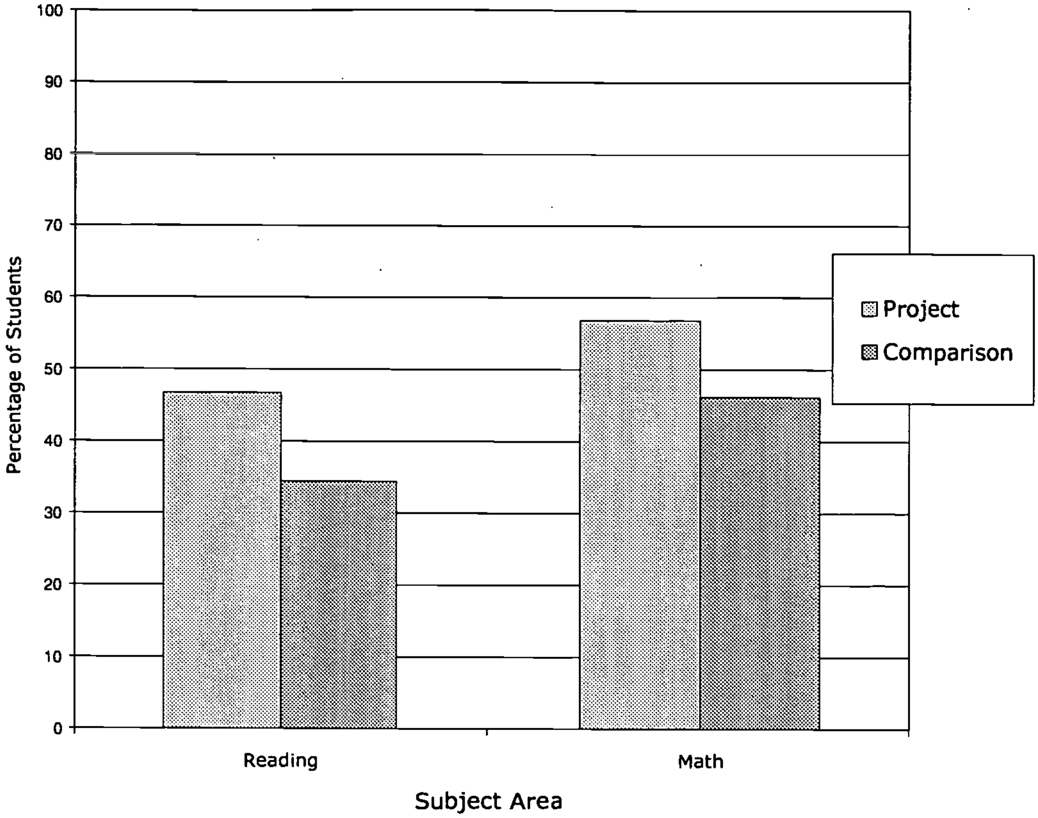


Figure 4. Comparison group model

Comparison Group Model—An Example

The Jefferson Elementary School has an enrollment of 500 students in kindergarten through grade five. The school has a very diverse student population with 35 percent minority students. Approximately 60 percent of the students are in the free or reduced-price lunch program. Jefferson has just adopted a comprehensive school reform model—Reading Enhancement—for schoolwide implementation. A school leadership team is formed to oversee the school improvement effort.

The statewide assessment program conducts testing of students in grades three and five in two core subject areas—reading and mathematics. The assessment takes place in April each year. The school also participates in districtwide writing assessment with grade five students in April each year.

The school leadership team wants to know whether with the implementation of the school reform model students at Jefferson are performing better than students in comparable schools (e.g., schools with similar demographic characteristics). The team chooses to use the comparison group model to conduct an impact evaluation of the school reform model. To the extent feasible and appropriate, the evaluation will take advantage of existing data available from the statewide assessment program to assess impact.

Step #1

The school leadership team, following extensive discussions with school staff, parents, and members from the community, decides to look at student performance in four areas: reading, mathematics, writing, and attendance. Even though the school reform model is focused on reading, the school and the community feel that it is important to look at other success indicators for the entire school.

Step #2

Most of the pertinent data will come from the statewide assessment program, including student achievement in reading and mathematics. Writing assessment data (for grade five only) will come from the district office. The only data collection instrument that needs to be created is a data form to provide summary data on student attendance—number of days absent per school year.

Step #3

In consultation with district-level staff, the leadership team identifies two schools in the district that are demographically similar to Jefferson. In School A, about 58 percent of the students are in the free or reduced-price lunch program.

In School B, the percentage is 62. Both schools have a diverse student population, with 35 percent minority students. School A has an enrollment of 400 students in kindergarten through grade five. School B has an enrollment of 600 students in the same grade span. Neither School A nor School B is implementing a comprehensive school reform program. The Jefferson leadership team decides that both School A and School B will be used as comparison schools in the evaluation.

Step #4

Student achievement data are obtained—electronically when feasible—from the statewide assessment program for grades three and five. There are approximately 70 or fewer students in each of these grades at Jefferson and the comparison schools. Data are obtained for all the students. No sampling is needed or desired. In addition, writing assessment data are obtained for all students in grade five. Attendance data are collected from school attendance records for all students in grades three and five. No sampling procedures are used.

Step #5

As in past years, after receiving training in test administration from state-level staff as part of the statewide assessment process, the classroom teachers administer the criterion-referenced tests in reading and mathematics to students in grades three and five in April at both Jefferson and the comparison schools. The tests, which have been aligned with the state content standards, consist of multiple-choice items and a few open-ended items. The tests are scored by a vendor and the results provided to the school and district as well as the state department of education.

In addition, the writing assessment is conducted with students in grade five following procedures established by the district office. Student attendance data are obtained from school records at Jefferson. For the comparison schools, attendance data are provided by the district office.

Advantages. This model has relatively strong scientific rigor, making it easier to attribute outcomes to the intervention. It is quite feasible when we can find naturally existing comparison groups (that is, student groups in a demographically similar school). In addition, it allows us to compare progress toward meeting common criteria (such as state standards).

Disadvantages. It is often difficult to find an appropriate comparison group. In addition, the selected groups may differ in important but unknown ways.

Another disadvantage is that data need to be collected for both intervention and comparison students, increasing the data collection burden and cost.

Step #6

A database is set up to store and manage all the data. The database contains statewide assessment data (reading and mathematics), districtwide writing assessment data, as well as attendance data for both Jefferson and the comparison schools. The data are analyzed to provide percentages of students (grades three and five) who meet the state standards or benchmarks in reading, mathematics, and writing for both Jefferson and the comparison schools. For comparison purposes, results for the two comparison schools are combined to provide a single percentage for each subject area. A difference in percentage points between Jefferson and the comparison schools provides an indication of impact.

In addition, an analysis is conducted on the mean differences of standard scores in reading and mathematics as well as ratings in writing assessment between Jefferson and the comparison schools (combined). A t test is performed to determine the statistical significance of each mean difference. A significant difference indicates that the intervention has, with a certain statistical probability, made a real difference in student performance. In addition, for each grade and each subject area, an effect size is calculated to assess the magnitude or educational significance of the difference.

Attendance data are analyzed to provide an average (mean or median) number of days absent per school year for Jefferson students and their counterparts at the comparison schools.

Step #7

Results of the analysis are provided in reader-friendly data displays (e.g., bar charts and line graphs) and easy-to-understand narratives. They are shared and discussed with stakeholder groups, including school staff, site council, parents, and members of the community.

Step #8

The results are provided to the district office and the state department of education to determine whether adequate progress has been made by the school. In addition, a meeting is held with the school leadership team, other key school staff, parents, and community members for an indepth review of the data to explore plausible reasons for the findings and to develop recommendations and an action plan for continuous improvement.

Regression Model

Using a statistical procedure called regression analysis, the model predicts or projects what things would have been like had there been no intervention (Fetler & Carlson, 1985; Yap, Estes, & Hansen, 1979; Yap, Estes, & Nickel, 1988; Yap, [September] 1980). The projection can take into account a range of factors that may have an influence on the outcomes, including demographics and current status of affairs. Typically, baseline status and relevant demographic variables are included in the regression equation. Other things being equal, the difference between actual outcomes and predicted outcomes is attributable as impact of the intervention. In the example shown on the following page, the project students as a group (or individually) scored higher on the state assessment than the level predicted by the regression equation.

The regression model is in many ways analogous to the baby growth chart one sees in a doctor's office. Based on such relevant information as a child's age, gender, and what is known about normal growth, the chart provides an expectation of the child's height and weight. Similarly, based on a student's grade level, current achievement status, and other relevant variables, the regression model provides an expectation on the student's achievement growth in core subject areas.

Implementation Steps. Important steps in implementing the regression model include the following:

1. Decide what outcomes you want to look at

2. Select or develop instruments to collect the pertinent data

3. Identify and obtain data needed to develop a regression equation

4. Develop a regression equation to predict outcomes

5. Decide whether sampling is desired

6. Administer the instruments to target groups

7. Analyze and interpret the evaluation data

8. Report findings to stakeholder groups

9. Use evaluation data for accountability and program improvement

The following example illustrates the use of the regression model to assess the impact of a school reform model.

Regression Model—An Example

The Jefferson Elementary School has an enrollment of 500 students in kindergarten through grade five. The school has a very diverse student population with 35 percent minority students. Approximately 60 percent of the students are in the free or reduced-price lunch program. Jefferson has just adopted a comprehensive school reform model—Reading Enhancement—for schoolwide implementation. A school leadership team is formed to oversee the school improvement effort.

The statewide assessment program conducts testing of students in grades three and five in two core subject areas—reading and mathematics. The assessment takes place in April each year. The school also participates in districtwide writing assessment with grade five students in April each year.

The school leadership team wants to know whether student performance is improving with the implementation of the school reform model. Given the

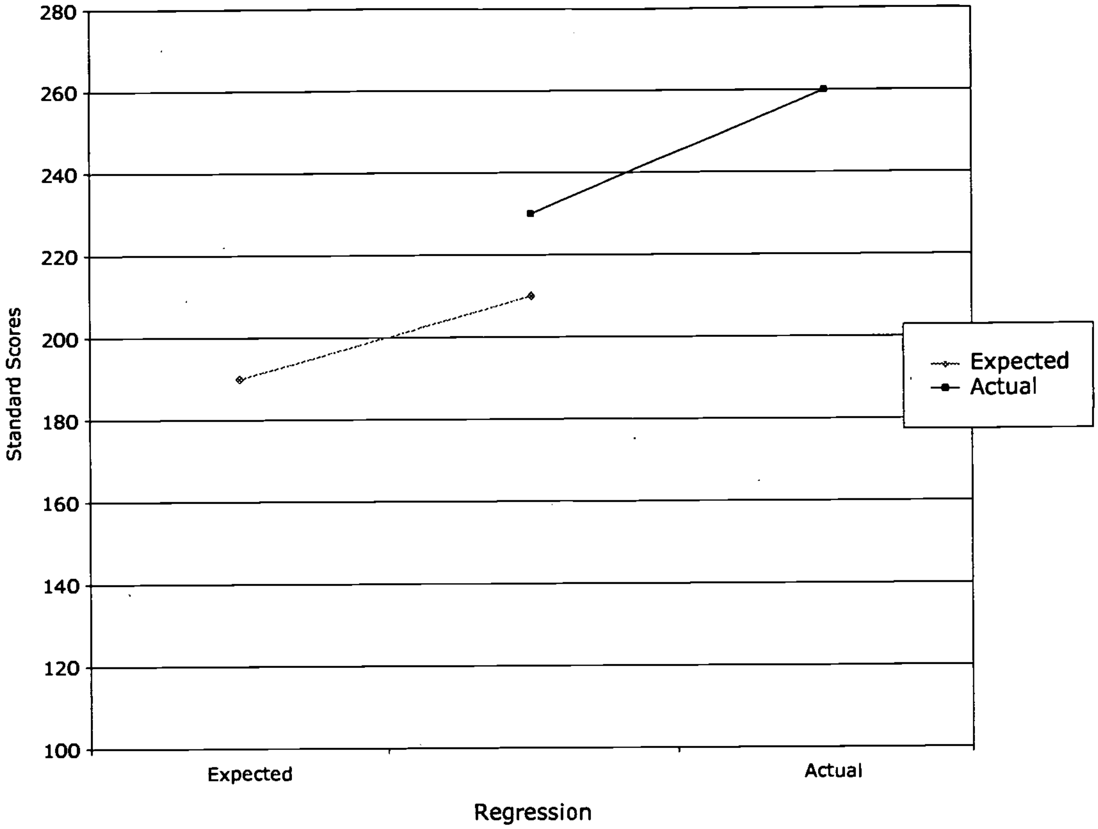


Figure 5. Regression model

intervention, are students performing as well as expected? The team chooses to use the regression model to conduct an impact evaluation of the school reform model. To the extent feasible and appropriate, the evaluation will take advantage of existing data available from the statewide assessment program.

Step #1

The school leadership team, following extensive discussions with school staff, parents, and members from the community, decides to look at student performance in two core subject areas: reading and mathematics. Even though the school reform model is focused on reading, the school and the community feel that it is important to look at student performance in mathematics as well.

Step #2

Most of the pertinent data will come from the statewide assessment program, including student achievement in reading and mathematics. Relevant school-level demographic data, including percent of students in free or reduced-price lunch program and percent of minority students, will also be obtained from the statewide assessment data system. No new data collection instruments are needed.

Step #3

Working with an external evaluator, the school leadership team decides that three types of data will be included in the regression equation: student achievement in reading and mathematics (for preceding school year and current school year), percent of students in free or reduced-price lunch program, and percent of minority students. In the regression analysis, the predictor variables will include student achievement for the preceding school year, percent of students in free or reduced-

Advantages. The models can have a high level of scientific rigor if the projection includes all of the pertinent factors. It takes advantage of existing data and does not require data collection from a control or comparison group. It statistically controls for extraneous factors affecting outcomes, making it possible to attribute program effects.

Disadvantages. The feasibility of the model depends in large measure on the availability of sufficient archival data—data that already exist—on the pertinent variables. The model requires statistical skills that may not exist among school staff. In addition, because it is essentially a statistical procedure, the model can often be misused.

price lunch program, and percent of minority students. The criterion or outcome variable is student achievement for the current school year. The evaluator will develop separate regression equations for reading and mathematics, using schools as units of analysis. The analysis will use school average scores—for grades three and five—instead of individual student scores.

To achieve sufficient reliability, the leadership team feels that the regression equation should be based on all 120 elementary schools in the state with grades three and five. All necessary data are obtained—electronically when feasible—from the statewide assessment data system.

Step #4

Using an appropriate data analysis package (e.g., SPSS or Excel), the external evaluator develops two separate regression equations to predict third-grade achievement—one for reading and one for mathematics. For each subject area, the equation predicts student achievement on the basis of achievement status for the preceding year, the percent of students in free or reduced-price lunch program, and the percent of minority students. For each of the schools included in the regression equation, the average scale score for third-graders is used as a measure of student achievement.

The evaluator develops similar regression equations to predict fifth-grade achievement.

Step #5

All elementary schools in the state with grades three and five are included in the regression equation. No sampling procedures are used.

Step #6

As in past years, after receiving training in test administration from state-level staff as part of the statewide assessment process, the classroom teachers administer the criterion-referenced tests in reading and mathematics to students in grades three and five in April. The tests, which have been aligned with the state content standards, consist of multiple-choice items and a few open-ended items. The tests are scored by a vendor and the results provided to the school and district as well as the state department of education.

The criterion-referenced tests provide a standard score in reading and mathematics for each student.

Step #7

The regression equations provide **predicted** achievement levels (i.e., average standard scores) for third- and fifth-graders in reading and mathematics. The predicted average standard scores are compared with the **actual** average standard scores of third- and fifth-graders at Jefferson. The difference is interpreted as an indication of impact of the school reform model on student performance.

The regression analysis identifies a cluster of four schools that most closely resemble Jefferson with respect to demographics. The average standard scores of these schools (combined) are compared with the average standard scores of Jefferson for third- and fifth-graders, respectively. The difference provides another indication of impact. A *t* test is performed to determine the statistical significance of each mean difference. A significant difference indicates that the intervention has, with certain statistical probability, made a real difference in student performance. In addition, for each grade and each subject area, an effect size is calculated to assess the magnitude or educational significance of the difference.

In addition, for each grade and each subject area, the percentage of students meeting state standards and benchmarks at Jefferson are compared with the percentage of students meeting standards and benchmarks at the four demographically similar schools. The difference provides yet another indication of impact.

Step #8

Results of the analysis are provided in reader-friendly data displays (e.g., bar charts and line graphs) and easy-to-understand narratives. They are shared and discussed with stakeholder groups, including school staff, site council, parents, and members of the community.

Step #9

The results are provided to the district office and the state department of education to determine whether adequate progress has been made by the school. In addition, a meeting is held with the school leadership team, other key school staff, parents, and community members for an indepth review of the data to explore plausible reasons for the findings and to develop recommendations and an action plan for continuous improvement.

Control Group Model

This is a true experimental design. Properly implemented, it requires random assignment of students to the intervention and control groups. Random assignment ensures the comparability or equivalence of the two groups in all pertinent respects other than the intervention itself (The Joint Committee on Standards for Educational Evaluation, 1994). Any difference between the two groups with respect to outcomes is therefore directly attributable to program effects. In the example shown on the following page, higher percentages of project students meet state standards in reading and mathematics in comparison with their control counterparts.

Implementation Steps. Important steps in implementing the control group model include the following:

1. Decide what outcomes you want to look at

2. Select or develop instruments to collect the pertinent data

3. Set up a control group through random assignment of students or other entities of interest

4. Decide whether sampling is desired

5. Administer the instruments to both project and control groups

6. Analyze and interpret the evaluation data

7. Report findings to stakeholder groups

8. Use evaluation data for accountability and program improvement

The following example illustrates the use of the control group model to assess the impact of a school reform model.

Control Group Model—An Example

The Jefferson Elementary School has an enrollment of 500 students in kindergarten through grade five. The school has a very diverse student population with 35 percent minority students. Approximately 60 percent of the students are in the free or reduced-price lunch program. Jefferson has just adopted a comprehensive school reform model—Reading Enhancement—for schoolwide implementation. A school leadership team is formed to oversee the school improvement effort.

The statewide assessment program conducts testing of students in grades three and five in two core subject areas—reading and mathematics. The assessment takes place in April each year. The school also participates in districtwide writing assessment with grade five students in April each year.

The school leadership team wants to know whether with the implementation of the school reform model students at Jefferson are performing better than they would have without the intervention. The team wants to use an evaluation model with a high level of scientific rigor—the control group model—to assess the impact of the school reform effort. To the extent feasible and appropriate, the evaluation will take advantage of existing data available from the statewide assessment program.

Step #1

The school leadership team, following extensive discussions with school staff, parents, and members from the community, decides to look at student performance in four areas: reading, mathematics, writing, and attendance. Even though the school reform model is focused on reading, the school and the community feel that it is important to look at other success indicators for the entire school.

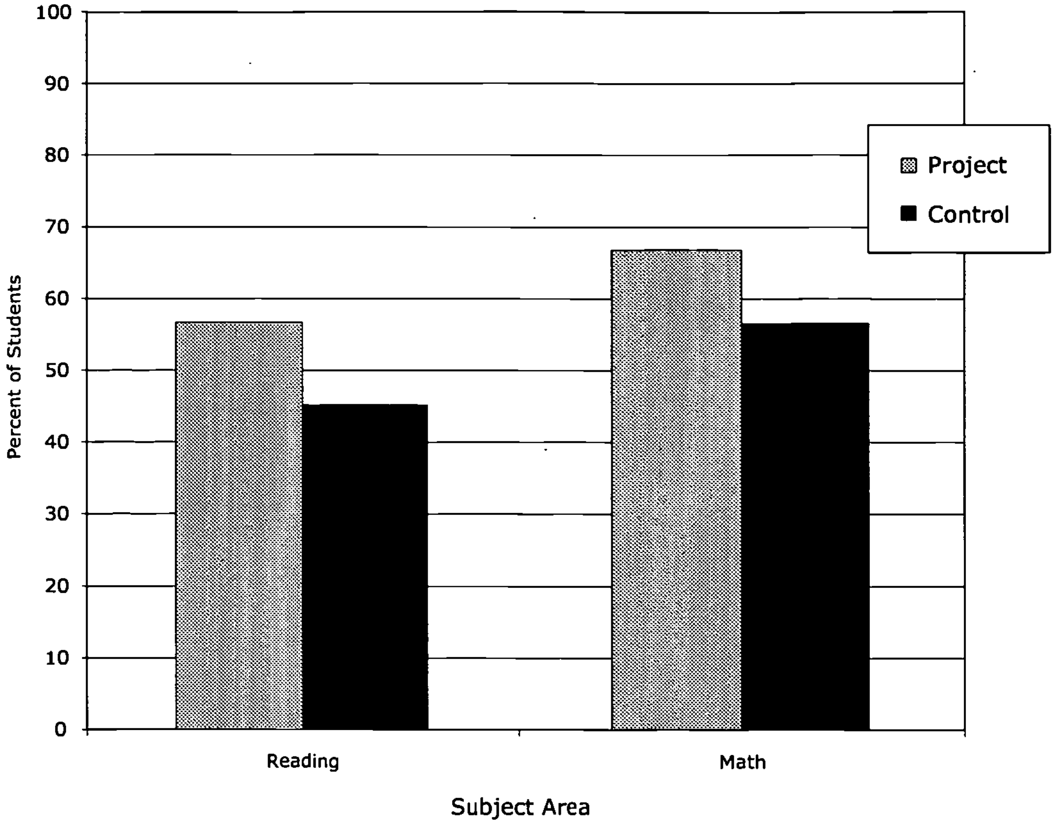


Figure 6. Control group model

Step #2

Most of the pertinent data will come from the statewide assessment program, including student achievement in reading and mathematics. Writing assessment data (for grade five only) will come from the district office. The only data collection instrument that needs to be created is a data form to provide summary data on student attendance—number of days absent per school year.

Step #3

Jefferson has three classes of third-graders and three classes of fifth-graders. Prior to the adoption of the school reform model, the leadership team works closely with the school administration and the teaching staff to reach a decision that, for the current school year, two of the three classes in each grade will participate in Reading Enhancement. The other class will serve as the control group. Furthermore, the school administration is able to persuade parents to allow students to be randomly assigned to the classes.

Step #4

Student achievement data are obtained—electronically when feasible—from the statewide assessment program for grades three and five. There are approximately 70 or fewer students in each of these grades at Jefferson. Data are obtained for all the students. No sampling is needed or desired. In addition, writing assessment data are obtained for all students in grade five. Attendance data are collected from school attendance records for all students in grades three and five. No sampling procedures are used.

Step #5

As in past years, after receiving training in test administration from state-level staff as part of the statewide assessment process, the classroom teachers administer the criterion-referenced tests in reading and mathemat-

Advantages. The model has a high level of scientific rigor. It provides the strongest basis for attributing the detected difference to the intervention. It has the potential of ruling out all extraneous factors that might have contributed to the outcomes.

Disadvantages. The model is probably the least feasible to implement, particularly in a school setting. It is almost never feasible to randomly assign students to the intervention and control groups. The process can be very disruptive. Another disadvantage is that it requires data collection for both the intervention and control groups, increasing data burden and cost.

ics to students in grades three and five in April. The tests, which have been aligned with the state content standards, consist of multiple-choice items and a few open-ended items. All students—those participating in Reading Enhancement and the control group students—take the tests. The tests are scored by a vendor and the results provided to the school and district as well as the state department of education.

In addition, the writing assessment is conducted with students in grade five following procedures established by the district office. Student attendance data are obtained at the end of the school year from school records.

Step #6

A database is set up to store and manage all the data. The database contains statewide assessment data (reading and mathematics), districtwide writing assessment data, as well as attendance data for both project and control students. The data are analyzed to provide percentages of students who meet the state standards or benchmarks in reading, mathematics, and writing—separately for project and control students. A difference in percentage points between the two groups provides an indication of impact.

In addition, an analysis is conducted on the mean differences of standard scores in reading and mathematics as well as ratings in writing assessment between project and control students. A *t* test is performed to determine the statistical significance of each mean difference. A significant difference indicates that the intervention has, with certain statistical probability, made a real difference in student performance. In addition, for each grade and each subject area, an effect size is calculated to assess the magnitude or educational significance of the difference.

Attendance data are analyzed to provide an average (mean or median) number of days absent during the school year for project and control students.

Step #7

Results of the analysis are provided in reader-friendly data displays (e.g., bar charts and line graphs) and easy-to-understand narratives. They are shared and discussed with stakeholder groups, including school staff, site council, parents, and members of the community.

Step #8

The results are provided to the district office and the state department of education to determine whether adequate progress has been made by the school. In addition, a meeting is held with the school leadership team, other key school staff, parents, and community members for an indepth review of the data to explore plausible reasons for the findings and to develop recommendations and an action plan for continuous improvement.

Table 1 provides a summary of the models along with their respective advantages and disadvantages.

Table 1. Advantages and Disadvantages of Evaluation Models

Model	Description	Advantages	Disadvantages
Pretest-Posttest	This model provides an expectation of program outcomes based on the current status.	■ Highly feasible in a school setting ■ Shows growth against baseline ■ Shows patterns and trends if conducted longitudinally ■ Can assess relative or absolute growth	■ May lack rigor—difficult to attribute effects to program ■ Difficult to control extraneous factors
Comparison Group	This model provides an expectation of program outcomes based on a comparable group.	■ Relatively strong scientific rigor ■ Can attribute effects to program ■ Can compare progress toward meeting common criteria (e.g., state standards)	■ May be difficult to find a comparable group ■ Selected groups may differ in some important but unknown ways ■ Increased data collection burden
Regression	This model uses a statistical method to predict or project program outcomes.	■ Relative strong scientific rigor ■ Can statistically control for extraneous factors affecting outcomes ■ Does not require existing control or comparison groups	■ Feasibility depends on availability of sufficient archival data ■ Model can be misused ■ Statistical expertise generally not available among existing school/district staff
Control Group	This model provides an expectation of program outcomes based on what happens in an equivalent or control group.	■ Has the strongest scientific rigor with random assignment of students to intervention ■ Can statistically control for extraneous factors affecting outcomes ■ Can attribute effects to program	■ Can compare progress toward meeting common criteria (e.g., state standards) ■ May be difficult, if not impossible, to find an equivalent group ■ Random assignment is typically not feasible in a school setting ■ Increased data collection burden

The Evaluation Process

Regardless of which model is used, the evaluation process consists of a series of critical steps, including the following:

Questions

1. What questions do we want to address?
2. What do we want to look at? What indicators and measures do we use?
3. How do we collect the data?
4. How do we analyze the data?
5. How do we interpret the data? What are the data telling us?
6. How do we use data to improve the program? What follow-up actions should be taken?
7. Are follow-up actions making a difference?

These steps are interrelated. Each is further discussed below.

Questions To Address

For impact evaluation, the overall question is whether and in what ways the intervention has made a difference for students, teachers, and the school as a whole. However, under this overall question, a host of more specific questions may be addressed by the evaluation. Examples include:

- How is the school and/or district administration providing support for the school reform effort?
- In what ways are teachers changing and improving their instructional practice?
- In what ways are students improving their performance?

Evaluation questions can be framed with even greater specificity as follows:

- Does the school reform effort result in an increased percentage of third-grade students meeting state benchmarks in reading and mathematics?
- Does the school reform effort result in an increased percentage of teachers participating in professional development activities?
- Does the school reform effort result in improved student attendance?
- Does the school reform effort result in a decreased number of discipline problems?

Some of these questions may have come directly from the stakeholders. Others may be based on stated program goals and objectives. Yet others may address specific program performance indicators. It is important that all key stakeholders are involved in making the decision on what questions the evaluation should address.

It is important that all key stakeholders are involved in making the decision on what questions the evaluation should address.

Choosing Indicators and Measures

Once the evaluation questions are formulated, it is normally an easy step to decide what indicators (such as reading achievement or student attendance) and measures (scores on specific tests, for instance) we need to look at. As discussed earlier, these indicators can exist at various levels: school/district administration, teachers, and students. For example, if the question has to do with the percent of students meeting state standards, then indicators may include student achievement in various academic areas (such as reading/language arts and/or mathematics).

Typically, indicators include student performance scores on the following measures:

- Norm-referenced tests
- Criterion-referenced tests
- Performance-based assessments

Norm-referenced tests (NRTs) are the most widely used standardized assessment tool in the United States. Their primary purpose is to provide a general portrayal of student performance in comparison with a norm group. A norm-referenced test typically consists of multiple-choice items in the areas of reading, language arts, mathematics, science, and social studies. Typically developed by a commercial publisher, NRTs provide such normative scores as percentiles, stanines, normal curve equivalents (NCEs), grade equivalents, and scale scores. These metrics are highly efficient for sorting and screening purposes, but are limited in indicating what students know and can do at a particular grade level.

Criterion-referenced tests (CRTs) are developed to assess the attainment of specific knowledge and skills. The test items, in a multiple-choice or an open-ended format, are constructed to measure a particular skill or instructional objective (for example, sight vocabulary, reading fluency, recognition of the central theme of a story, addition with two-digit numbers, basic algebraic concepts). In most cases, a cut score or mastery score is established to determine whether a student has mastered a specific skill. In this sense, assessments based on state standards or benchmarks are a form of criterion-referenced testing. Many states are using the services of commercial publishers to create their standards-based assessment systems.

It is important to recognize that in addition to academic subjects, other indicators may also be pertinent, including the following:

- Attendance
- Dropout rates
- Discipline referrals
- Violence

Performance-based assessments (PBAs) are created to provide students with opportunities to apply or demonstrate specific knowledge or skills in a particular content area. While a consensus has yet to emerge on a precise definition of performance-based assessments, such assessment devices generally require the student to create a response to an open-ended question. Examples include a short written answer, a writing sample, an exhibition, and a portfolio. The response is typically scored or rated according to a set of specific criteria described in a scoring guide or rubric. The best-developed and most widely used performance-based assessment is traits-based writing assessment. Student writing samples are typically rated on a six-point scale for such traits as ideas, organization, word choice, voice, and conventions. PBAs allow teachers to incorporate assessment as an integral part of instruction.

Norm-referenced measures are not consistent with the notion that all students will attain a particular level of knowledge and skills.

Also typically, these assessment devices cover the following academic areas:

- Reading/language arts
- Mathematics
- Writing
- Science
- Social studies

In standards-based school reform, it is probably more appropriate to look at indicators that are standards-based rather than norm-referenced. Most states have both content and student performance standards that address the question of what students should know and be able to do at various benchmark points. In this context, a critically important indicator is the percentage of students meeting the state standards. Because they measure students against one another, rather than against an external standard, norm-referenced measures are not consistent with the notion that *all* students will attain a particular level of knowledge and skills.

In addition to student outcomes, the evaluation may also look at indicators at the school and teacher levels. At the school level, we may want to find out whether and how the school administration is supporting the reform effort. Changes in policy and practice can occur in the following areas:

- Release time for teachers to plan improvement activities
- Reallocation of time and resources for professional development
- Acquiring external technical assistance to enhance staff capacity

At the teacher level, the evaluation may look at the following:

- Incidence of collegial learning
- Use of effective teaching practice
- Redesigning the curriculum
- Use of assessment information to improve instruction

Collecting Data

Several decisions need to be made here. For example, key decisions need to be made in the following areas:

- Which evaluation model is the most appropriate for addressing the questions?
- What instruments should be used to collect the data?
- What are the data sources?
- Is sampling necessary or desired?
- Should we use multiple measures?

Model Selection. Quite often, the evaluation question itself would suggest which evaluation model may be the most appropriate. For example, if we are interested in knowing not only whether the percent of students meeting state standards is increasing but also whether the increase is greater than a comparable group, then the comparison group model is appropriate. On the other hand, if we are interested in knowing only whether the school is improving over time, then a pretest-posttest model may suffice.

A model is seldom, if ever, entirely valid or invalid. Some models are generally more valid than others. There are other criteria schools should consider in choosing a particular model.

First, we need to consider the purpose of the evaluation. When an evaluation is conducted for formative purposes (e.g., for program modification and refinement), the ability to make a causal link may be less important than when it is conducted for high-stakes, summative purposes (e.g., for program continuation). A less rigorous model may be adequate for exploratory, formative investigations.

Second, we need to consider feasibility. Generally, less vigorous models are easier to implement than more rigorous models. For example, a true experimental design with random assignment of students to experimental and control groups is typically not feasible in the regular school setting. The use of naturally existing comparison groups, while less rigorous, is more feasible. Other factors related to feasibility include the intrusiveness of data collection procedures as well as staff time and expertise for data collection and analysis. For example, when teachers and school administrators serve as data collectors, data collection methods need to be explicit and relatively straightforward.

> The evaluator must weigh the importance and usefulness of the information against the resources needed to collect and analyze the data.

Third, cost is always an important consideration. Generally, the more rigorous models are more expensive than their less rigorous counterparts. The evaluator must weigh the importance and usefulness of the information against the resources needed to collect and analyze the data. The model selected should provide benefits commensurate with the costs it incurs.

Instrument Selection. Depending on the nature of the specific indicators you are looking at, various instruments may be appropriate for data collection. For example, if the indicators have to do with academic achievement, some sort of tests for assessment devices will be required for data collection. If the indicators deal with teaching practice, a different set of instruments will be used to collect the relevant data. Such instruments may include interview protocols, observation schedules, and/or focus-group meetings. Like the evaluation models, each data collection method has its advantages and disadvantages.

Researchers and evaluators have developed a variety of data collection methods, including:

- Document review
- Questionnaire survey
- Interview
- Focus group
- Observation
- Assessment of student achievement

Some methods are better suited for the collection of certain types of data. Each has advantages and disadvantages in terms of costs and other practical and technical considerations (such as ease of use, accuracy, reliability, and validity). For example, there is no best way to conduct interviews. Your approach will depend on the practical considerations of getting the work done during the specified time period. Using a focus group—which is essentially a group interview—is more efficient than one-on-one interviews. However, people often give different answers in groups than they do individually. They may feel freer to express personal views in a private interview. At the same time, group conversations can draw out deeper insights as participants listen to what others are saying. Both approaches have value. Schools must weigh pros and cons against program goals.

For both focus groups and interviews, the evaluator should work from a written interview guide that lists the questions and also provides space where the interviewer can record answers. Good interview questions should be open-ended questions written in a clear, simple, conversational style.

If your data collection plan calls for classroom observations, the evaluator needs to develop a guide that describes what he or she is looking for in the classroom. For example, the observer may be asked to look for ways the inservice training has changed classroom practice. Or she may be asked to note whether the teacher is using certain program materials. During the visit itself, the evaluator should avoid disrupting the classroom activity. It is best if the evaluator sits in an unobtrusive place and uses the guide to focus on the relevant classroom actions.

The Data Collection Matrix on the next page summarizes the advantages and disadvantages of each method.

Table 2. Data Collection Matrix

Method	Focus	Advantages	Disadvantages
Document Review	■ Nature and level of school reform activities ■ Incidence of events of interest ■ Existing student achievement information	■ Data already exist ■ Low cost ■ Typically unobtrusive ■ Relatively unbiased	■ Lack of quality control ■ Validity and reliability may be unknown ■ Can be limited in scope
Interview	■ School staff/parent/student perceptions ■ School staff/parent satisfaction ■ Improvement suggestions ■ Degree of implementation ■ Anticipated and unanticipated outcomes	■ Indepth information ■ Quality control ■ High response rate ■ Opportunity to probe	■ Relatively costly ■ Needs trained data collectors ■ Data can be biased ■ May require careful sampling
Survey	■ School staff/parent/student perceptions ■ School staff/parent/student satisfaction ■ Improvement suggestions ■ Degree of implementation ■ Anticipated and unanticipated outcomes	■ Relatively low cost ■ Can include structured and open-ended information ■ Relative ease of administration ■ Can cover a large number of respondents	■ Response rate often a problem ■ Needs careful sampling ■ Data can be biased ■ Open-ended data may be difficult to analyze

Table 2. Data Collection Matrix (continued)

Method	Focus	Advantages	Disadvantages
Focus Group	■ School staff/parent/student perceptions ■ School staff/parent/student satisfaction ■ Implementation issues ■ Improvement suggestions ■ Degree of implementation ■ Anticipated and unanticipated outcomes	■ Indepth information on program implementation and outcomes ■ Relatively free of response rate problems ■ Interactive discussion among stakeholders	■ Relatively high cost ■ Needs trained facilitators ■ May be difficult to achieve appropriate representation in recruitment of participants ■ Group dynamics can bias discussion
Observation	■ Program implementation ■ Classroom activities ■ Instructional practices ■ School climate	■ Increased objectivity and authenticity of data ■ Can provide contextual information	■ Needs trained observers ■ Relatively high cost ■ Can be obtrusive ■ Often just a snapshot of program implementation ■ May not reflect typical reality
Assessment	■ Student performance in cognitive and affective domains	■ Objective data often with known reliability and validity ■ Can be low cost (standardized testing) ■ Can include large samples of students ■ Provides a generally accepted portrayal of schooling outcomes	■ May provide a limited and narrow picture of student performance ■ Can be high cost (performance-based assessments) ■ May need careful sampling

Case studies are not listed as a data collection method because they typically employ some or all of the data collection methods under conditions specified in a fieldwork plan. A well-designed case study not only provides a rich documentation of program implementation and outcomes but can often help make a logical connection between program activities and the desired outcomes.

Data Sources. Various sources exist from which the evaluator may collect the pertinent data. Archival sources consist of existing documents from which a wide array of data (such as student assessment data, attendance, and discipline referrals) may be available. The primary data sources will probably be people who are participating in the school reform effort, including students, teachers, school administrators, parents, and community members. Typically, survey and interview data on program implementation and outcomes will come from teachers, school administrators, parents, and community members. Achievement data will be gathered from students.

While each data source can provide valuable information on the selected indicators, care should be taken in deciding which data source may be best for which type of information.

While each data source can provide valuable information on the selected indicators, care should be taken in deciding which data source may be best for which type of information. For example, data on teacher professional development can come from teachers or school administrators. Generally, teachers will be a better data source in this case because they have firsthand knowledge of the staff development activity and can provide a more valid and accurate picture of what took place and its potential impact. Similarly, in some cases, teachers' self-reports on instructional practice may be less accurate than data obtained from onsite observation by a trained observer.

In addition, many data sources can be strengthened by some preparatory work. For example, a good explanation of the purpose of the evaluation, clear and concise instructions for completing a written survey, and a well-developed focus group guide can all enhance the validity of the data. Making sure that students know the purpose of a particular assessment and have adequate test-taking skills can also increase the validity and accuracy of the assessment results.

Multiple Measures. In many instances, it is unlikely that a single measure will adequately assess the extent to which a program objective is attained, especially when the objective entails complex and multifaceted knowledge and skills on the part of students or teachers. In such cases, the use of multiple measures and approaches can enhance the validity, reliability, equity, and utility of the data as well as decisions about teaching and learning. Multiple measures should be used to capitalize on the strengths of each data collection method. For example, survey data on changed practices at the classroom level can be supplemented with onsite observation data to enhance validity. Similarly, the validity of student performance data is enhanced when such data are gathered with different approaches and formats, including criterion-referenced tests, multiple-choice tests, writing samples, completion of tasks and projects, and portfolios of student work.

The use of multiple measures and approaches can enhance the validity, reliability, equity, and utility of the data as well as decisions about teaching and learning.

Sampling. Sampling can reduce data collection cost as well as burden on respondents. Matrix sampling, for example, allows a selected sample of the target population (for example, teachers or students) to respond to a selected sample of test or survey items. It reduces the amount of time and other resources for data collection in comparison with a study that requires the participation of all members of the target group. On the down side, sampling reduces the amount of information available for individual students and teachers, and may make it difficult to disaggregate data.

Sampling can reduce data collection cost as well as burden on respondents.

Sampling units can be individuals (such as students or teachers), grade levels, schools, districts, or even states in a large-scale study. A simple random sample of individual students will consist of students randomly selected from the entire school, district, state, or nation. Similarly, a simple random sample of schools will consist of schools randomly selected from the district, state, or nation.

The most efficient sampling method (with the smallest sampling error) is stratified random sampling (Sudman, 1976). For example, within a school, you can first randomly sample grade levels and then randomly select students within each grade level selected. The stratification factors can be any variables that may potentially affect the outcomes, including grade level, gender, ethnic group, and poverty status.

Data Quality. Selecting and using an appropriate evaluation model, instruments, data sources, sampling methods, and multiple measures will help ensure that high quality data are collected for the evaluation. Several criteria can be used to assess data quality, including validity, reliability, accuracy, and utility.

Validity is the most important consideration. The selected instrument, whether it is a norm-referenced test, a criterion-referenced test, or performance-based assessment, should measure what it is supposed to measure. For example, a test consisting of only multiple-choice items is not likely to provide valid information on students' higher order thinking skills. The selected instrument should have construct validity in the sense that it measures concepts and skills that are the targets of instruction. For standards-based assessment, the instrument should be aligned with state content standards as well as classroom instruction.

Reliability refers to the consistency of assessment results. For example, a test should provide very similar, if not identical, results if it is given to the same students twice over a short period of time (e.g., a week or two). When this is the case, the test is said to have high test-retest reliability. In addition, the items in the test should "hang together" in the sense that they measure the same skills and knowledge as indicated by an internal consistency measure. In writing assessment, reliability means that two or more trained raters using the same scoring rubrics should provide highly similar, if not identical, ratings for the same writing samples.

Accuracy means that the assessment results are relatively free of measurement or sampling errors. These errors can come from poor test administration, use of inappropriate sampling procedures, and/or inadequate attention to scoring rubrics. Error sources can be minimized by developing clearly written instructions for test administration and scoring. When measurement errors are known to exist, they should be taken into account in data interpretation.

Finally, high quality data should also be user-friendly. This is particularly important when data are intended to be used by school staff to improve instruction or the entire program. It is critical that the data be meaningful to teachers and school administrators if they are expected to use the data for improvement purposes. Involving school staff and parents in designing data collection activities can go a long way to enhancing data **utility**—that the data will be used as intended.

Data Collection Schedule. Depending on the impact questions being addressed and the evaluation model used, data need to be collected at appropriate times during the school year. In many cases, the evaluator may be able to take advantage of data collection procedures that have already been put in place (e.g., a statewide assessment system). In other cases, the evaluator may be able to use archival data (i.e., data that have already been collected). In any case, it is helpful to conduct a "data audit" to find out any and all existing data that can be used to address the evaluation ques-

It is helpful to conduct a "data audit" to find out any and all existing data that can be used to address the evaluation questions before initiating new data collection activities.

tions before initiating new data collection activities.

In general, evaluation data should be collected repeatedly over time to demonstrate patterns and trends of student performance. For example, in the pretest-posttest model, data should be collected for at least two time points (e.g., at the beginning and end of a school year). It is, however, helpful to continue to collect data for additional time points on a regular basis (e.g., fall-spring, spring-spring, fall-fall, or some other annual cycles) over several years. This allows us to show performance trends and patterns as well as the sustained effects of the intervention. For the other models, longitudinal data are similarly desirable. The chart on this page provides examples of schedules for collecting student performance data for each evaluation model.

The schedules on this page are examples only and should be modified to take advantage of existing data collection activities. For example, statewide assessment, which often provides much of the needed student performance data, may occur in March (or some other time of the school year) instead of April. In that case, March or another month of the school year will become the pretest and/or posttest date.

Also, student performance data may be collected more frequently than fall to spring or once a year for instructional improvement purposes. Many comprehensive school reform models require the collection of

Evaluation Models & Data Collection Schedules

Pretest-Posttest

Option A: Fall-Spring
September (Pretest)
April (Posttest)

Option B: Annual (Spring-Spring)
April–April (Pretest-Posttest)

Note. In this model, data are collected from project students only.

Comparison Group

Option A: Pretest-Posttest
September (Pretest)
April (Posttest)

Option B: Posttest Only
April (Posttest)

Note. In this model, data are collected from both project and comparison group students.

Regression

September/October—Collection of demographic and other relevant contextual data (e.g., free or reduced-price lunch status and pretest scores)

April—Collection of posttest data

Note. In this model, data are collected from a larger population of students of which the project students may be a part (e.g., districtwide or statewide student population).

Control Group

Option A: Pretest-Posttest
September (Pretest)
April (Posttest)

Option B: Posttest Only
April (Posttest)

Note. In this model, data are collected from both project and control group students.

assessment data on an ongoing basis (e.g., every eight weeks). Such data can and should be used for instruction planning as well as for impact evaluation.

Data Management. There is a wide range of software packages that the evaluator can use to manage the evaluation data. For example, the evaluator can set up a database with SPSS, Access, or Excel. Each requires a different level of technical expertise. For a relatively small school or district, Excel—the simplest of the three programs—should work well as a database software. For schools or districts with larger student enrollments, SPSS or Access may be more efficient. For all software packages, the user manual typically provides instructions for setting up and managing a database.

Regardless of which software is used, the database should have the following capabilities:

- Include student achievement data on core subject areas (e.g., reading/language arts and mathematics)
- Include individual student demographic information (e.g., gender, ethnicity, migrant status, language proficiency status, disability status, economically disadvantaged status)
- Include data on other contextual variables (e.g., attendance, teacher-student ratio, instruction, discipline, and violence)
- Track student performance over time (e.g., several years)
- Aggregate and disaggregate data (e.g., for total student population and various subgroups)
- Include procedures for data analysis using both descriptive and inferential statistics

To keep the database current and usable, it is critically important that a staff member be designated to maintain the database once it is set up. This includes clear and specific procedures for data entry in a timely manner and periodic checks on data quality. In many cases, student performance data can be extracted or exported electronically from other databases (e.g., statewide assessment data systems) into the evaluation database.

Analyzing the Data

The most commonly used statistics include the following:

Frequency Count. A frequency count provides an enumeration of activities, things, or people that have certain pre-specified characteristics. Examples include:

- Number of teachers who participated in professional development activities
- Number of minutes of class time devoted to reading
- Number of students meeting state standards in reading
- Number of days absent for the average student per school year

Frequency counts can often be categorized (e.g., 0, 1-5, 6-10, more than 10) in data analysis.

Percentage. A percentage tells us the proportion of activities, things, or people that have certain characteristics within the total sample. Examples include:

- Percent of students in grade four meeting reading benchmarks
- Percent of minority students at a school
- Percent of students in a school district living in poverty
- Percent of teachers in a state participating in professional development activities

Percentage is probably the most commonly used statistic to show the current status as well as growth over time. For example, a school or district may set a goal to increase the proportion of students meeting state benchmarks by 5 percent each year.

Mode. The mode is the most frequently occurring number in a data set. For example, in a writing assessment, if the most frequent rating is 3 (on a 6-point scale), then the mode rating is 3. The mode tells us what is the most typical case. In some instances, it gives us a better picture of what is going on than other statistics (e.g., the mean).

Median. The median is the middle or 50th percentile score. This is a good statistic to use to represent the average when the score distribution is nowhere near normal. For example, in looking at attendance data, the median gives us a much better picture than the mean if a few students were absent for a huge portion of the school year. Unlike the mean, the median is much less affected by a few outlying or extreme scores.

Mean. The mean is the most commonly used statistic to represent the average in research and evaluation studies. It is derived by dividing the sum by the total number of units (e.g., teachers or students) included in the summation. It tells us what the average teacher or student is like with respect to performance. The mean has mathematical properties that make it appropriate to use with many statistical procedures (e.g., test of statistical significance of a difference between two groups).

Standard Deviation. Standard deviation shows the spread of a score distribution—the larger the standard deviation, the wider the spread. In survey data, it indicates the extent to which the respondents provided similar responses or ratings. When the respondents provided the same or highly similar responses, the standard deviation of their responses will be small. A large standard deviation, on the other hand, suggests less agreement among the respondents.

It is important, however, that conclusions and recommendations regarding program implementation and outcomes be based on patterns and trends of results rather than episodic differences.

In most instances, data analysis will be straightforward, using such descriptive statistics as frequency counts, averages, and percentages. It is important, however, that conclusions and recommendations regarding program implementation and outcomes be based on patterns and trends of results rather than episodic differences that may represent little more than measurement errors or random fluctuation over time.

Data disaggregation can help identify areas in which a program is succeeding and areas in which improvement is needed. It can also identify areas where equity is an issue.

Data analysis is facilitated if the project has clear and measurable goals and objectives (Yap, 1997). For example, if an objective of the project is to increase the percentage of third-graders meeting state reading benchmarks, then it is a relatively simple matter to compute the number and percent of these students who met the benchmarks.

In some cases, you may want to use "inferential" statistics to analyze the data, especially if the evaluation has a high-stakes purpose, such as program funding. This is where you want to be sure that the detected differences (positive or negative) are not a result of random fluctuation. A variety of statistical procedures (such as a *t* test for differences between two groups or analysis of variance among three or more groups) are available to assess the statistical significance of a detected difference. If such technical expertise is not available among the school staff, external help can be obtained to perform the analysis.

In addition, data should be disaggregated whenever possible. For example, data can be broken down by gender, ethnic group, school locale (urban and rural), and student type (economically disadvantaged, limited-English proficient, migrant, disabled, and so forth).

Schools and districts with Title I projects are required to disaggregate assessment data by:

- Major racial and ethnic group
- Gender
- English proficiency status
- Migrant status
- Disability status
- Economically disadvantaged status

Schools and districts must report the disaggregated data unless the number of students in any group is too small to provide statistically sound information or would reveal the identity of individual students. The most recent guidance (U.S. Department of Education, 1999, p. 49) from the U.S. Department of Education suggests that disaggregated data for subgroups of fewer than 10 students are probably not statistically sound and should not be reported.

While schools are not required to report disaggregated data for small samples, such data can and should be used for purposes of instructional or program improvement. In addition, there are ways of increasing the sample size to make the disaggregated results more representative. For example, student achievement data can be combined over time (e.g., for two or more consecutive years) or across grade levels for the same subject area to create a larger student sample for data disaggregation.

Data disaggregation can help identify areas in which a program is succeeding and areas in which improvement is needed. It can also identify areas where equity is an issue. For example, disaggregation can serve as protection against "creaming"—a deliberate or unconscious attempt on the part of program staff to achieve better results by working only with more advantaged or promising students. "Creaming" is not only discriminatory, it also undermines the integrity of standards-based reform.

Interpreting the Data

This is where we ask the question: What are the data telling us? Contrary to a common belief, data do not usually speak for themselves. The results must be interpreted in an appropriate context. For this reason, interpretation is best conducted as a collaborative activity between the evaluator and project staff. For example, differences in student performance over time can be a result of random fluctuation. The evaluator with statistical expertise can help decide whether that is the case or whether the difference is statistically related to the intervention. Project staff, however, are generally in a better position to discuss the meaning of the difference and its implications for teaching and learning.

A wide array of test scores are used to measure student performance, including the following:

Raw Scores. A raw score is simply the number of test items that a student answered correctly. For example, in a 60-item test, if the student responded correctly to 45 items, then her raw score is 45. A raw score, which cannot exceed the total number of items in a test, has no inherent meaning.

Percent Correct. This is the proportion of test items that a student answered correctly. In the above example, where the student responded correctly to 45 of the 60 items in a test, her percent correct score is 75—she responded correctly to 75 percent of the items included in the test. It is important that we do not confuse percent correct scores with percentile scores.

Ratings. Ratings are typically provided in performance assessments. For example, writing samples are often rated by trained raters on a 6-point scale based on clearly defined rubrics or scoring guides. A student may receive a rating of 4, for example, for her writing sample. Ratings can be provided for the writing sample as a whole (holistic scoring) or for each of the traits of interest (e.g., ideas, voice, organization, conventions).

Percentiles. Percentiles, a norm-referenced metric, indicate the percent of students in the norming sample—typically a nationally representative sample—who scored below a certain score. For example, if a student scores at the 60th percentile, it means that 60 percent of the students in the norming sample scored below her score. Roughly speaking, she scores better than 60 percent of the students included in the norming sample. Percentile scores range from 1 to 99.

Interpretation is best conducted as a collaborative activity between the evaluator and project staff.

Quartiles. Quartiles are cutpoints in a particular score distribution. Technically, there are three quartiles—at the 25th, 50th, and 75th percentiles—which divide the distribution into four equal portions. For example, the top quartile consists of students who score at or above the 75th percentile. The bottom quartile consists of students who score at or below the 25th percentile.

Stanines. Stanines are a nine-point scale created and used by the U.S. Army during World War II to screen out feeble-minded recruits. It has since enjoyed widespread use in education for screening and selection purposes. Stanines provide an efficient way of sorting students into nine categories. Quite often, students are grouped in low (1-3), middle (4-6), and high (7-9) stanines.

Normal Curve Equivalents. Normal curve equivalents (NCEs) were originally created for use in the evaluation of Title I projects. The metric is closely related to the percentile scale. Like percentiles, NCE scores range from 1 to 99. In fact, the two scales coincide at three points: 1st, 50th, and 99th percentiles. Psychometrically, the critical difference between the two metrics is that NCEs form an equal-interval scale whereas percentiles do not. Being an equal-interval scale, NCEs are appropriate for use in statistical calculations (e.g., in the computation of means and standard deviations).

Grade Equivalents. Grade equivalent scores form a longitudinal scale to assess the mastery of skills and knowledge from kindergarten through the 12th grade. The school year is conceptually divided into 10 learning months, the three months in summer being considered as one learning month. Grade equivalents typically range from K to 12. Thus, a grade equivalent score of 2.5 means that the student scores at a learning level of second grade and five months. Grade equivalents are derived from a complicated scaling process, which can often create confusion or result in misunderstanding and misuse of the metric. Suppose a second-grade student taking a second-grade test obtains a grade equivalent score of 3.0. What does that mean? It means that had the average third-grade student taken the second-grade test at the beginning of the school year, she would have gotten the same score as the second-grade student. Conversely, suppose a third-grade student taking a third-grade test obtains a grade equivalent score of 2.0. It means that had the average second-grade student taken the third-grade test at the beginning of the school year, she would have gotten the same score as the third-grade student. To add to the complexity, grade equivalents are typically based on statistical projections rather than test scores from real students. Thus, the meaning of "falling behind grade" or "scoring above grade" is not as straightforward as it might seem.

Standard Scores. Standard scores form a longitudinal scale to assess the mastery of skills and knowledge from kindergarten through the 12th grade. Derived from a sophisticated scaling process, standard scores link the various test levels in a battery of norm-referenced or criterion-referenced tests into a single scale. Normally, a student in a lower grade is expected to have a lower standard score than a student in a higher grade. As a student moves on to higher grades, her score is expected to increase. Standard scores can serve as cut-scores for various levels of proficiency in a core subject area (e.g., partially proficient, proficient, and advanced in reading or mathematics). In this sense, they are particularly useful in standards-based assessments. Typically a three-digit number, standard scores have other names such as scale scores or expanded standard scores.

Two other considerations are critically important in interpreting test scores. First, not all test scores have equal intervals. For example, percentiles and most grade equivalent scores are not equal-interval scales. They are not suitable for use in the calculation of various statistical indices (e.g., mean and standard deviation). This is because a unit on the scale may have different meaning and importance relative to other units, depending on where it is on the scale. For example, on the percentile scale, the units are narrower or tighter in the middle range than those at the high or low end. The NCE scale, on the other hand, consists of units of equal size along the entire scale.

Second, some test scores are status scores in the sense that they show the achievement status of a student or a group of students relative to other students. Percentiles, NCEs, and stanines are examples of status scores. On the other hand, longitudinal scores indicate where a student or a group of students is on a continuum of skills or content knowledge. Standard scores and grade equivalents are examples of longitudinal scores.

The following matrix provides a classification of the commonly used test scores along the two dimensions.

Equal/non Equal-Interval

Status Scores

Equal-Interval	Non Equal-Interval*
■ Stanines	■ Percentiles
■ Normal curve equivalents	■ Quartiles
■ Percent correct	■ Raw scores
■ Ratings	

Longitudinal Scores

Equal-Interval	Non Equal-Interval*
■ Standard scores	■ Grade equivalents

*Not appropriate for direct statistical computation (e.g., calculation of mean and standard deviation). Strictly speaking, raw scores are not an equal-interval scale, even though they are often used in statistical computation.

Evaluators commonly say that a difference is "significant" or "not significant." Typically, they are referring to the statistical significance of a difference between the experimental or project students and the control/comparison students. A significant difference in this sense merely means that it is unlikely that the detected difference is a result of random fluctuation. For example, when a difference is said to be significant at the .05 level—a conventional level of significance—it means that the difference can be a result of random fluctuation only about 5 percent of the time. To the extent that 5 percent is considered a low probability, one may conclude that the difference is probably not due to random fluctuation and, in that sense, is a real difference.

However, a "real" difference may be small or large. It does not tell us anything about the practical or educational value of the difference. The value or practical importance of the difference is essentially a judgment call, to be determined by the key stakeholders participating in the intervention. Evaluators have come up with some rules of thumb to assess the practical importance of a difference. A common rule is that if the difference is more than one-third of the standard deviation, it may be considered as having some practical importance. The normal curve equivalent (NCE) scores, for instance, have a standard deviation of approximately 21. A difference of 7 or more NCEs may therefore be considered to have practical importance.

Project staff, with intimate knowledge of program implementation, can help provide a more complete explanation of the outcomes. For example, demographic changes or a sudden influx of transient students can significantly affect student outcomes. Such extenuating circumstances need to be considered if data interpretation is to have credibility with project staff who are expected to use the evaluation results to improve program implementation and outcomes.

Data interpretation is greatly facilitated if the project has set up measurable goals and objectives or has developed performance indicators that are readily assessable. Objectives or performance indicators that incorporate a standard or criterion make it easy to conclude whether the objective has been met. For instance, if an objective requires 60 percent of the third-graders to meet state benchmarks, it is a relatively easy task to decide if the objective is attained.

Using Data for Program Improvement

Results of impact evaluation can serve a dual purpose: accountability and program improvement (Kushman & Yap, 1999). Just like findings from program implementation evaluation, results of impact evaluation should also be useful to the project staff. While we need to know if the program is achieving the goals and objectives it set out to achieve, it is also important that project staff be able to use the impact information to plan follow-up actions to further strengthen the program.

Objectives or performance indicators that incorporate a standard or criterion make it easy to conclude whether the objective has been met.

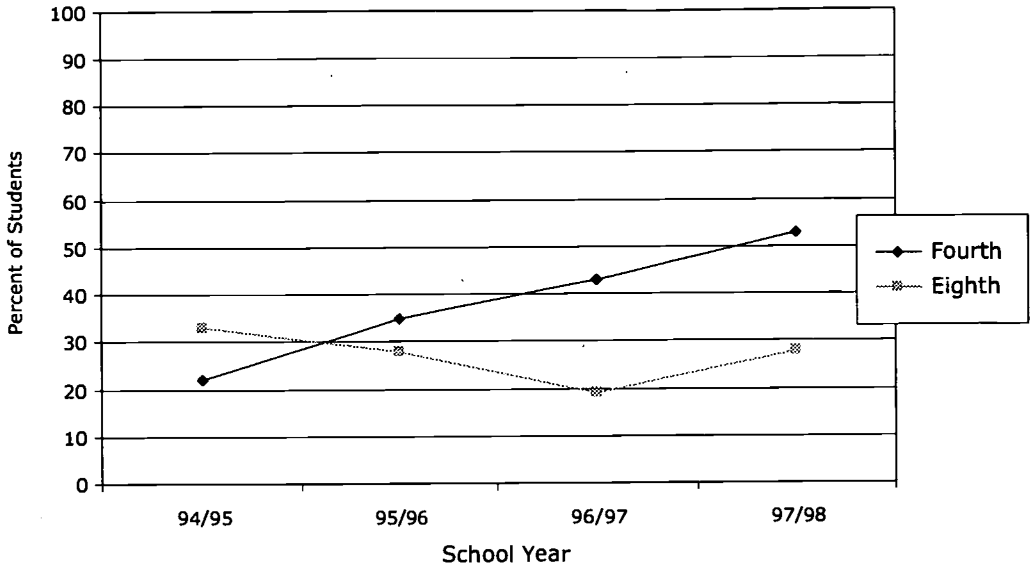


Figure 7. Percent of students meeting mathematics benchmarks

Like data interpretation, data use is best conducted as a collaborative activity between the evaluator and project staff. The evaluator can present the data and findings in a way that is understandable and useful to project staff, who can then develop plans for program modification and refinement. A good way to do this is for the evaluator and project staff to engage in an interactive discussion on outcomes. For example, the evaluator can prepare the impact data in a graphical format as above:

It is also important that project staff be able to use the impact information to plan follow-up actions to further strengthen the program.

In this example, the project staff will be asked to develop a set of narratives, using their own words, to describe what the data are telling them. This will be followed by discussion and clarification until a consensus or agreement is reached on what the data say and/or imply. An action plan will then be developed to implement follow-up activities. In the above example, there is clearly a need to re-examine and strengthen the eighth-grade mathematics curriculum.

The action plan may consist of the adoption or adaptation of a new comprehensive school improvement model or the development of a home-grown approach to school improvement. It may seek to expand professional development of school staff.

The activities should be research-based, challenging, and doable.

The action plan should have a time line and should identify individuals responsible for carrying out the planned activities. Like any program elements, the activities should be research-based, challenging, and doable. For example, if the corrective action calls for further professional development, then the plan should be based on the principles of effective practice in professional development, including:

- Activities are based on, and reflect, the best available research and practice
- Activities are ongoing, intensive, and sustained
- Content has direct application in practice
- Goals are developed with input from participants
- Goals are part of a long-term school improvement plan
- There is a formative (implementation) and summative (impact) evaluation process
- Key stakeholders are involved in both the evaluation and refinement of the professional development activities
- There is understanding among stakeholders of how professional development fits in the larger, overall school improvement plan

Monitoring Follow-Up Actions

The implementation of the follow-up action plan needs to be monitored and evaluated. Particular attention should be focused on the intent of the corrective action. For example, if the correction consists of increased professional development, then implementation evaluation during the following year should include professional development as a focus. Data should be collected to indicate whether professional development activities have increased (compared with the preceding year) and to assess the quality of such activities.

> The impact of the corrective action should be evaluated like other program components.

The impact of the corrective action should be evaluated like other program components. This makes program evaluation, both implementation and impact, an integral part of the school improvement cycle—a process for continuous improvement.

Resources

Bernhardt, V.L. (1998). *Data analysis for comprehensive schoolwide improvement.* Larchmont, NY: Eye on Education.

This book presents practical tools to help educators effectively gather, analyze, interpret, and use data to make better decisions for comprehensive schoolwide improvement. Written for non-statisticians, the book shows the reader how to collect and use a variety of data such as demographics, attendance/enrollment, and assessment data.

Holcomb, E.L. (1999). *Getting excited about data: How to combine people, passion, and proof.* Newbury Park, CA: Corwin Press.

This practical manual answers questions about what data to collect, how to analyze data, and how to interpret and use the data for schoolwide improvement.

Levesque, K., Bradby, D., Rossi, K., & Teitelbaum, P. (1998). *At your fingertips: Using everyday data to improve schools.* Berkeley, CA: MPR Associates, Berkeley, CA: National Center for Research in Vocational Education, & Arlington, VA: American Association of School Administrators.

This workbook is designed to help educators use a variety of data to better manage, monitor, and improve schools. The workbook is structured to help teams and individuals develop performance indicator systems that can be used to identify strengths and weaknesses, and to develop educational strategies to meet educational goals.

References

Blum, R.E., Yap, K.O., & Butler, J.A. (1991). *Onward to Excellence impact study.* Portland, OR: Northwest Regional Educational Laboratory.

Fetler, M.E., & Carlson, D.C. (1985). Identification of exemplary schools on a large scale. In G. Austin & H. Garber (Eds.), *Research on exemplary schools* (pp. 83-96). Orlando, FL: Academic Press.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.

Kushman, J.W., & Yap, K.O. (1999). What makes the difference in school improvement? An impact study of Onward to Excellence in Mississippi schools. *Journal of Education for Students Placed at Risk, 4*(3), 277-298.

Messick, S. (1985). Progress toward standards as standards for process: A potential role for NAEP. *Educational Measurement: Issues and Practice, 4*(4), 16-19.

Sudman, S. (1976). *Applied sampling.* New York: Academic Press.

Tallmadge, G.K. (1982). An empirical assessment of norm-referenced evaluation methodology. *Journal of Educational Measurement, 19*(2), 97-112.

U.S. Department of Education. (1999). *Peer review guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act.* Washington, DC: Author.

Yap, K.O. (1980). Pretest-posttest correlation and the special regression model. In *American Statistical Association: 1980 proceedings of the social statistics section* (pp. 236-240). Washington, DC: American Statistical Association.

Yap, K.O. (1980, September). *Pretest-posttest variance differentials and the special regression model.* A paper presented at the annual meeting of the American Psychological Association, Montreal, Canada.

Yap, K.O. (1997). *Guidebook on developing performance indicators.* Portland, OR: Northwest Regional Educational Laboratory.

Yap, K.O., Estes, G.D., & Hansen, J.B. (1979, April). *Effects of data analysis methods and selection procedures in regression models.* A paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Yap, K.O., Estes, G.D., & Nickel, P.R. (1988). *A summative evaluation of the Kamehameha elementary education program as disseminated in Hawaii public schools.* Portland, OR: Northwest Regional Educational Laboratory.

Workshop Requirements

The following are general requirements for this training activity:

Audience: District and school-level evaluators and key project staff responsible for the evaluation of whole-school reform efforts.

Time: Two to four hours

Group size: 20 to 30 participants

Equipment: An overhead projector and Chart-pack paper

Materials: Transparencies, participant handouts, and a copy of guidebook (desired)

Objective: To build local capacity in evaluating whole-school reform efforts through an interactive presentation and discussion on impact evaluation.

Begin the discussion by stating the primary purpose of impact evaluation—to find out if the intervention (whole-school reform) has made a difference for schools, teachers and, most important, students.

Then use the transparencies to continue with the presentation and discussion. The presentation should be as interactive as possible. Since the audience is likely to consist of people with considerable experience and expertise with program evaluation, you should invite questions and comments from the audience as much as possible.

Depending on the type of audience you have and how detailed the presentation/ discussion needs to be, this session can last two to four hours. For district or school staff responsible for program evaluation, this can be made a work session in which the participants will complete the small-group activities as preplanning for their evaluation work.

Instructions for Impact Evaluation Transparencies

Each transparency is related to a part of the guidebook. You should familiarize yourself with the contents of the guidebook before you use the transparencies. The guidebook generally gives you a pretty good idea about what you should say when you show a particular transparency.

Transparency #1

Explain that there are many ways to find out if an intervention has made a difference. Each evaluation model uses a different method and rationale to determine what things would have been like had there been no intervention. The difference between actual and expected outcomes is a measure of program impact.

The models are also different in that the results they produce allow us to attribute, with differing degrees of confidence, the outcomes to the intervention. They also differ with respect to feasibility, cost, and obtrusiveness. Thus, each has advantages and disadvantages. Discuss the advantages and disadvantages. Refer to Pages 50 through 62 in the guidebook.

Generally speaking, the models are presented in order of scientific rigor. The pretest-posttest model is the least rigorous and the control group model—a true experimental design—is the most rigorous. In a layperson's perspective, one may say that the models answer the following questions:

Pretest-posttest model—
Are things getting better?

Comparison group model—
Are you making a difference?

Regression model—Are you doing better than expected?

Control group model—Are you really making a difference?

Transparency #2

Present the pretest-posttest model as one that is highly doable and reasonable when evaluation resources and expertise are limited. It measures outcomes at a minimum of two time points—pretest and posttest. However, it is best conducted with measures repeated at regular intervals, for example, each fall and spring or annually.

The assumption of this model is that, without the intervention, things at posttest time will be the same as they were at pretest time. Teachers will teach the same way and students will learn the same way. Any difference will, therefore, likely be a result of the intervention.

Briefly discuss the advantages and disadvantages of the model as discussed on Page 51 in the guidebook.

Explain that the best way to use the pretest-posttest model is not just to do a pretest and a posttest. Rather, it should be repeated over a long period of time—preferably over several years to show longitudinal patterns and trends. Even though this model does not provide a strong scientific basis for attributing impact to the intervention, a consistently positive trend can be compelling evidence that the program is working.

See pages 50-53 in the guidebook.

Transparency #3

Present the comparison group model as one with relatively strong scientific rigor. It is generally doable when the school can find an appropriate comparison group—a school or groups of students with characteristics similar to those of students in the intervention. At the very least, the two groups (or schools) should be demographically similar, including such factors as poverty level, percent of minority students, LEP population, and so on.

The assumption of this model is that, without the intervention, things (including the way teachers teach and the way students learn) will be very much alike, if not identical, at the project and comparison schools. Any difference found at the end of the intervention will, therefore, be attributable to the intervention.

One of the challenges of using this model is finding a comparison group that is similar to the intervention group in all relevant respects and one that is willing to participate in the necessary data collection activities. In some cases, some sort of incentive (such as a summary of findings of the study) may need to be provided to get such cooperation.

Briefly discuss the advantages and disadvantages of the model as described on Page 55 in the guidebook.

Transparency #4

Present the regression model as one that is of great interest to evaluators and researchers. While it is more doable in a school setting than people might think, it does require statistical skills not normally available among school staff. It is likely that some external assistance will be needed if this model is chosen.

The assumption of this model is that the regression procedure can provide a highly accurate prediction of what things would have been like in the absence of the intervention, especially when all relevant variables are accounted for in the equation. The difference (as shown in the transparency) between the predicted status and actual status at the end of the intervention period is attributable to the intervention.

The unit of measurement and analysis can be individual students, schools, or other entities of interest. For example, individual student scores can be used to establish the regression equation. This will probably be done by grade level. The procedure will then provide a predicted score for each student. On a larger scale, schools can be used as the unit in setting up the equation. In that case, school averages, for both student performance and demographics, will be used as the scores to be included in the regression equation. Again, this is best conducted by grade level. The equation will then provide a predicted score for each grade level for the school as a whole.

Briefly discuss the advantages and disadvantages of the model as described on Page 57 in the guidebook.

Transparency #5

Introduce the control group model as a true experimental design with the highest level of scientific rigor. Random assignment of students or other entities of interest to the intervention and control groups can potentially rule out all extraneous factors that may affect the outcomes, making it easy to attribute program impact.

The assumption of the model is that the project and control groups are truly equivalent in all relevant respects and, without the intervention, we would expect the same things to happen in both groups. If there is a difference at the end of the intervention period, that will be attributed to the intervention.

A challenge of the model is random assignment of students to project and control groups. This is rarely, if ever, feasible in an ordinary school setting. Randomly assigning larger entities (e.g., classes or schools) is sometimes more feasible. However, with larger entities, even random assignment may not result in truly equivalent groups.

The control group model, even though rarely feasible, serves as an ideal that schools can approximate to the extent possible. When this model is used, we can attribute the difference between the two groups, as shown in the transparency, to the intervention with a great deal of confidence.

Briefly discuss the advantages and disadvantages of the model as described on Page 60 in the guidebook.

Close the discussion of evaluation models by directing attention to Impact Evaluation Handout #1, which summarizes the advantages and disadvantages of each model.

Transparency #6

Walk the audience through the evaluation process, pointing out that steps are interactive and build on each other. It is important to point out that the project needs to set up measurable goals and objectives or performance indicators that can be assessed—those with some sort of standards or criteria built in.

Schools will probably want to look at outcomes at more than one level. For example, they might want to find out whether, as a result of the whole-school reform:

- School policy and practice have changed, particularly with respect to professional development and allocation of time and resources
- Instructional practice has changed
- Student performance patterns have changed

Students are the ultimate beneficiaries of school reform. It would be difficult to justify leaving out student outcomes in an impact evaluation of whole-school reform effort.

We need to look at the evaluation process from a cost-benefit perspective. For example, some models and data collection methods are more expensive or time-consuming than others. We need to make sure the expected benefits to the target groups (students, teachers, and schools) are commensurate with the cost incurred.

All of the steps, but especially the last three steps, in the process are best conducted as a collaborative effort between the evaluator and project or school staff. The evaluator can present the results and the project staff can bring their craft knowledge about the reform effort to help interpret the findings and to plan follow-up actions. Ultimately, only project staff—not the evaluator—can use evaluation data to improve the project.

Relevant contents are provided on Pages 63-78 of the guidebook.

Transparency #7

Explain that there are many ways of collecting evaluation data. Some are better suited for gathering certain types of data as discussed on Pages 65-72 in the guidebook. Some are more expensive than others. Each has advantages and disadvantages. Again, cost and benefits should be considered in data collection. Generally, more indepth information costs more and is more time-consuming to collect. For example, a written survey is usually less expensive than onsite observation but may provide only a very global picture of program implementation.

Briefly discuss each data collection method as described on Pages 67-68 in the guidebook.

At this point you may want to have the participants peruse the handout on data collection (Data Collection Matrix) and solicit comments and observations.

Transparency #8

Discuss data collection considerations as described on Pages 65-72 in the guidebook, reinforcing the notion that we want to collect data that are valid, reliable, and useful in the most cost-effective way.

Selecting the most appropriate model will give us the most valid data for the intended purpose.

Instruments must be valid, reliable, and cost-effective for the type of data we are collecting. For example, a written survey on teaching practice may be less expensive, but onsite observation (which is more expensive) can provide more accurate and useful data.

Some data sources may be more valid than others. As a general rule, we should go to the primary source. For example, if we want to know the extent to which teachers participate in professional development activities, the data source should be teachers, not a district administrator.

Sampling can reduce the cost of data collection. In some cases, sampling might even provide more accurate data where the response rate problem may be more serious.

Multiple measures give us a more comprehensive and therefore more accurate picture of program implementation and outcomes.

Discuss data quality, data collection schedule, and data management as described on Pages 70-72.

Transparency #9

Briefly discuss the difference between descriptive statistics and inferential statistics. In many cases, the use of descriptive statistics (e.g., frequency counts, percentages, averages) may suffice, especially when the evaluation does not have high stakes.

When it is necessary (such as in a high-stakes evaluation) to be sure that the impact is not a result of random fluctuation, inferential statistical procedures may be needed. In some cases, a *t* test to assess the statistical significance of the difference between the project and comparison group may be all that is needed. In others, analysis of variance or other more sophisticated procedures to detect a "real" difference may be necessary.

At this point, you may want to talk about different styles of data analysis. Data can be made to reveal the truth—which is what we are after—in various ways. For example, they can be squeezed, massaged, or brutally tortured to "confess" the truth as we see it.

You may also gently remind your audience that while some facets of the truth may readily ooze out of the data, other facets use data as a shield to hide their identity. Sophisticated, high-voltage statistical procedures may be needed to penetrate the shield to get to the whole truth. Even then, one should be reminded that there are lies, damned lies, and then statistics.

Back on a more serious note, you may want to discuss the difference between statistical significance and the practical importance of any detected difference. See Page 76 in the guidebook.

Evidence is more compelling when there is a consistent pattern or trend. For example, with the pretest-posttest model (which is generally less rigorous than the other models), if the student performance shows a consistently positive trend over multiple years, one may quite confidently say that something is going right with the intervention.

Whenever feasible, data should be disaggregated. Title I requires data to be broken down by gender, ethnicity, poverty, language, migrant status, and disability status. Disaggregated data provide us with a better understanding of how the intervention is working and can also reveal equity issues which may otherwise not surface. See Pages 73-74 in the guidebook.

Transparency #10

Explain that there are only a handful of statistical indices in common use. They are frequency count, percentage, mode, median, and mean/ standard deviation. Go over this quickly because most people in the audience probably already know these indices.

Frequency count tells us, for example, how many teachers participated in how many professional development activities, how many minutes of the class time were devoted to reading, how many students were absent for how many days, and so on. Frequency counts can often be categorized (0, 1-5, 6-10, more than 10) in data analysis.

Percentage tells us the proportion of teachers who participated in professional development activities, the proportion of students at various achievement levels (such as meeting state reading benchmarks), the proportion of students who dropped out, and so on. Percentage is probably the most commonly used statistic to show current status as well as growth over time. For example, a school or district may set a goal to increase the proportion of students meeting state benchmarks by 5 percent each year.

Technically, mode is the most frequently occurring number in a data set. For example, in a writing assessment, if the most frequent rating is 3 (on a 6-point scale) then the mode rating is 3. Mode tells us what is the most typical case. In some cases, it gives us a better picture of what is going on than the mean.

The median is the middle or 50th percentile score. This is a good statistic when the score distribution is nowhere near normal. For example, in looking at attendance data, the median gives us a much better picture than the mean if a few students were absent for a huge portion of the school year. The median is much less affected by a few outlying or extreme scores.

Mean and standard deviation are the most commonly used statistics in research and evaluation studies. The mean tells us the average—what the average teacher or student is like with respect to performance. For example, when we want to find out the difference between two groups (say, project and comparison groups) we compare the means for the two groups.

Standard deviation shows the spread of the score distribution—the larger the standard deviation, the wider the spread. In survey data, it indicates the extent to which the respondents provided similar responses or ratings. When the respondents provided the same or similar responses, the standard deviation of their responses will be small.

A larger standard deviation, on the other hand, suggests less agreement among the respondents.

Transparency #11

Show the transparency and go over the items quickly. Again, most people in the audience probably already know their test scores well.

Point out that ratings are typically used in performance-based assessment (e.g., writing assessment). Typically, the ratings are based on some well-developed scoring guide or rubrics. The ratings are usually single-digit numbers.

Point out that some test scores are not equal-interval scores, which means that they cannot be used in statistical calculation. For example, it is not appropriate to add and divide percentile scores to get an average. To get an average percentile, we should do the computation with Normal Curve Equivalent (NCE) scores and then convert the average NCE to a percentile score.

Strictly speaking, only stanines, NCEs, and standard scores are equal-interval scores.

Also, test scores can be divided into status (horizontal) and longitudinal (vertical) scores. The status scores (e.g., percentiles, quartiles, stanines, and NCEs) compare the performance of a group of students with that of their peers. Longitudinal scores (grade equivalents and standard scores) show or capture a vertical scale or continuum of knowledge or skills by grade level or a hierarchy of difficulty.

Transparency #12

Show Transparency #12 when you do Small-Group Activity #3. See Small-Group Activity #3 for details.

Transparency #13

Use Transparency #13 when you do Small-Group Activity #4. See Small-Group Activity #4 for details.

Impact Evaluation Small-Group Activities

Each small-group activity is designed to reinforce or stimulate the discussion on a particular topic or concept. They may be conducted before or after the discussion. If the activity is done before the discussion, the topic should be briefly introduced first. As a presenter, you should guide the participants through the activity and then lead an interactive discussion of the results of the groups' work, drawing from the contents of the guidebook as appropriate to reinforce and/or enrich the discussion.

The small-group activity can also be scheduled to follow a more detailed discussion of the topic. In this case, the activity provides a way for the participants to apply what they have learned in the presentation and discussion.

Small Group Activity #1 (20 minutes)

This activity can be conducted before or after your presentation on data collection (Impact Evaluation Transparencies #7 and #8). If it is conducted before the presentation, its purpose is to stimulate thinking about data collection issues. If it is done after the presentation, its purpose is to reinforce ideas and concepts covered in your presentation.

Divide the audience into groups of about five people. The group can consist of members of a school team or just participants selected by various means to form a group.

The task of the group is to complete the data collection form (Impact Evaluation Handout #3) to reinforce what they have discussed about data collection, including methods, data sources, and instruments. The small group should identify a recorder and/or reporter to share the results with the entire group when the activity is completed. Allow 15 minutes for the small groups to complete the task and five minutes to share. To save time, you may ask only two or three volunteer groups to share.

Refer the participants to parts of the guidebook that discuss evaluation models and data collection methods (for example, the data collection matrix).

As discussed in the guidebook, data collection methods can include document review, interview (in person or over the telephone), written survey, focus groups, observation, and assessment of student performance.

Data sources can include existing documents and people, including students, teachers, school administrators, parents, and community members.

Under "instrument," the small groups can provide generic labels (such as "teacher survey" or titles of existing instruments as in the measurement of student achievement by a statewide test).

At the end of the activity, you should briefly summarize the results and point out any common themes, patterns, or trends. If the concepts did not come up in the group discussion, you should briefly discuss the advantages and disadvantages of each data collection method with respect to validity, reliability, feasibility, cost, and data burden.

Small-Group Activity #2 (20 minutes)

This activity can be conducted prior to or following your presentation on data analysis (Impact Evaluation Transparencies #9, #10, and #11). If it is conducted before the presentation, its purpose is to stimulate thinking about data analysis issues. If it is done after the presentation, its purpose is to reinforce ideas and concepts covered in your presentation.

Divide the audience into small groups of about five people. The group can consist of members of a school team or just participants selected by various means to form a group.

The task of the group is to complete the data analysis form (Impact Evaluation Handout #4) to reinforce what they have discussed about data analysis, including the use of descriptive and inferential statistics.

The small group should identify a recorder and/or reporter to share the results with the entire group when the activity is completed. Allow 15 minutes for the small groups to complete the task and five minutes to share. To save time, you may ask only two or three volunteer groups to share.

Explain that under the column heading of type of data, we are talking about whether it would be survey data, interview data, observation data, student outcome data, or others.

Under data analysis method, members of the group should discuss whether they would compute frequencies, percentages, and/or averages. Would they set a standard or criterion? For example, would they want to see at least 50 percent of the teachers changing their instructional practice in accordance with what is specified in the school reform model? Would they look at student outcomes in addressing the evaluation question? How can they say instruction has improved unless students are learning better? Would they do any comparative analysis?

Would they be dealing with open-ended, qualitative data, such as descriptions of changes in practice? Would they just summarize the verbal data?

Small-Group Activity #3 (30 minutes)

Show Impact Evaluation Transparency #12 when you do Small-Group Activity #3.

Divide the audience into groups of about five people. The group can consist of members of a school team or just participants selected by various means to form a group.

The task for members of the group is to review the student outcome data (percent of students meeting state benchmarks) and to state in their own words what the data mean to them. Collectively, they are to develop three narratives or statements that indicate what the data say or imply. Typically, these narratives are then used as the basis for developing improvement plans.

The small group should identify a recorder and/or reporter to share the results with the entire group when the activity is completed. Allow 25 minutes for the small groups to complete the task and five minutes to share. To save time, you may ask only two or three volunteer groups to share.

At the end of the activity, you should briefly summarize the results and point out any common themes and findings.

Small-Group Activity #4 (30 minutes)

Use Impact Evaluation Transparency #13 when you do Small-Group Activity #4.

Divide the audience into groups of about five people. The group can consist of members of a school team or just participants selected by various means to form a group.

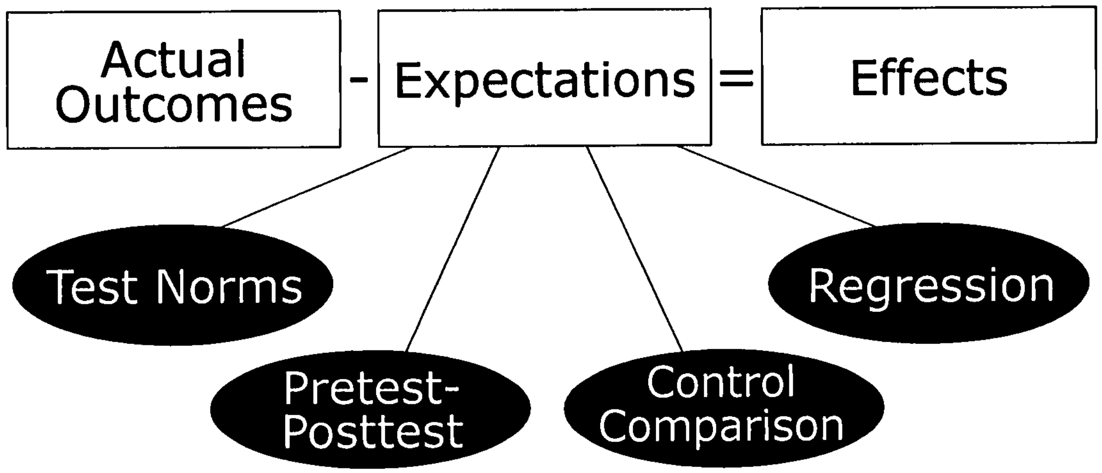
The task of the group is to review the student data displayed in a graph. The same data are provided in a tabular format for Small Group Activity #3. The group is to develop key findings based on the data in response to the evaluation question of whether student performance is improving over time.

Based on the key findings, the group will then decide what corrective action, if any, should be taken. The group will also decide who will be responsible for implementing the corrective action and when the action will be taken.

The small group should identify a recorder and/or reporter to share the results with the entire group when the activity is completed. Allow 25 minutes for the small groups to complete the task and five minutes to share. To save time, you may ask only two or three volunteer groups to share.

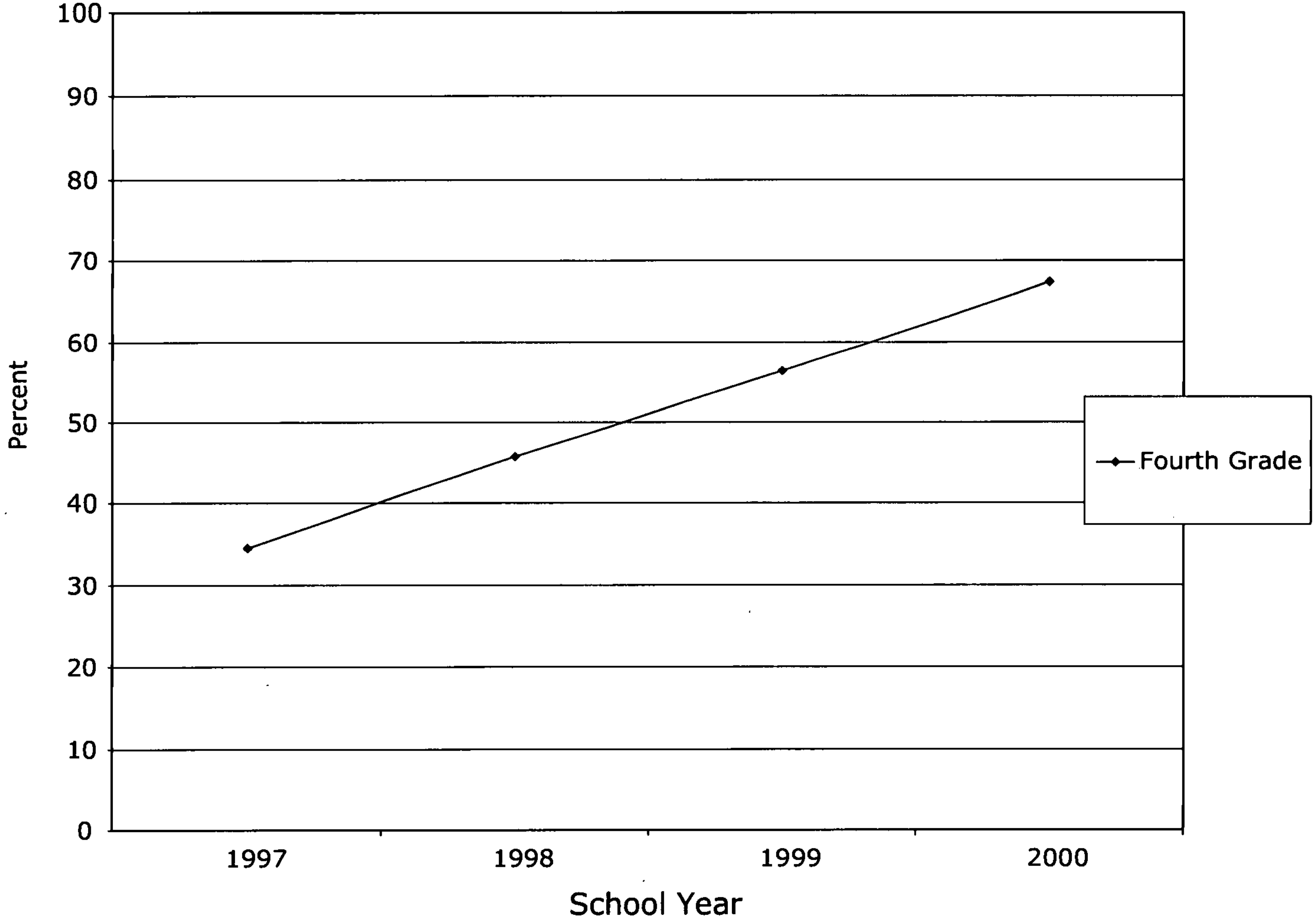
At the end of the activity, you should briefly summarize the results and point out any common themes, patterns, or trends. If none of the groups mentioned it, you should point out that the eighth-grade curriculum clearly needs to be examined and perhaps restructured.

Evaluation Model

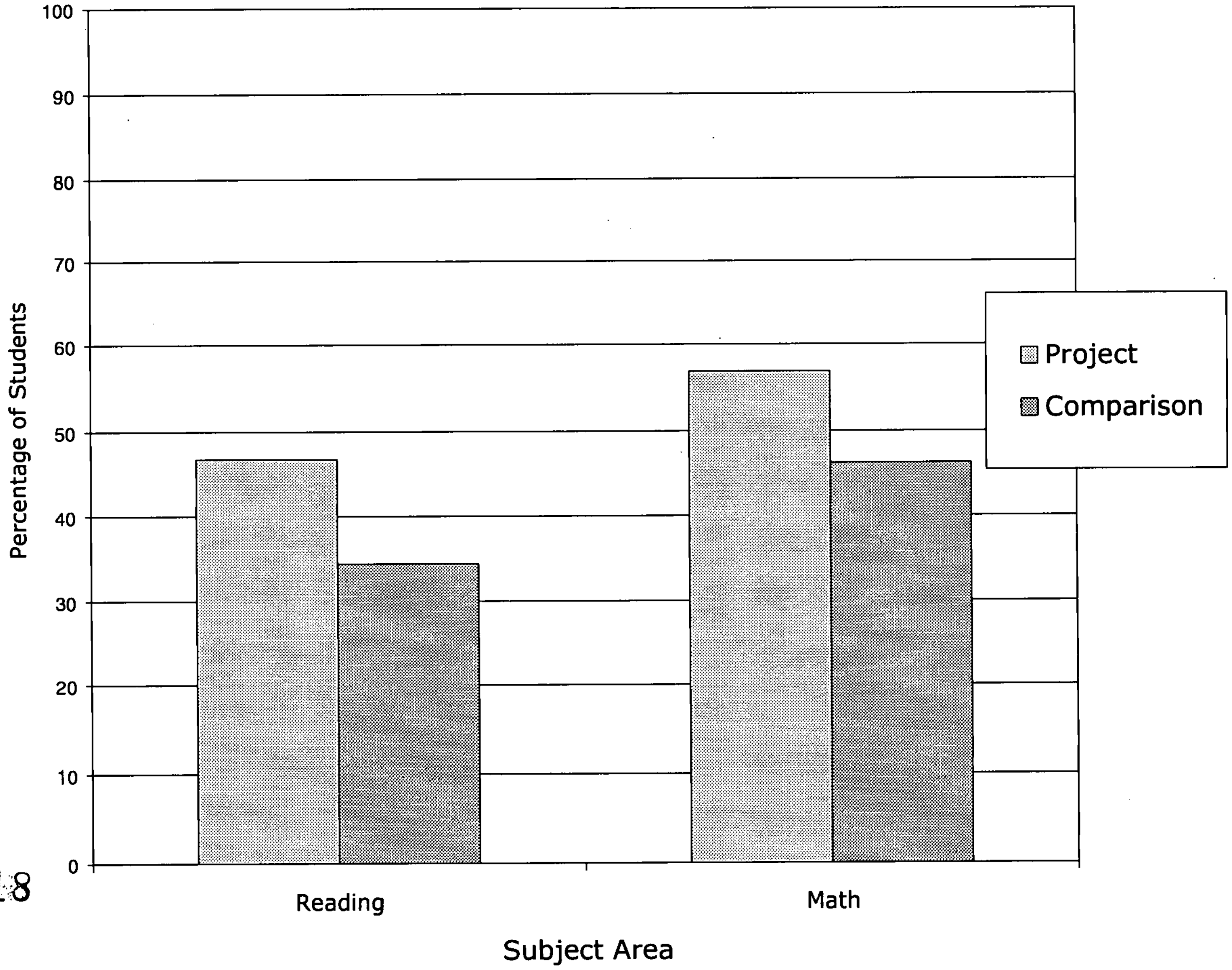


Actual Outcomes - Expectations = Effects

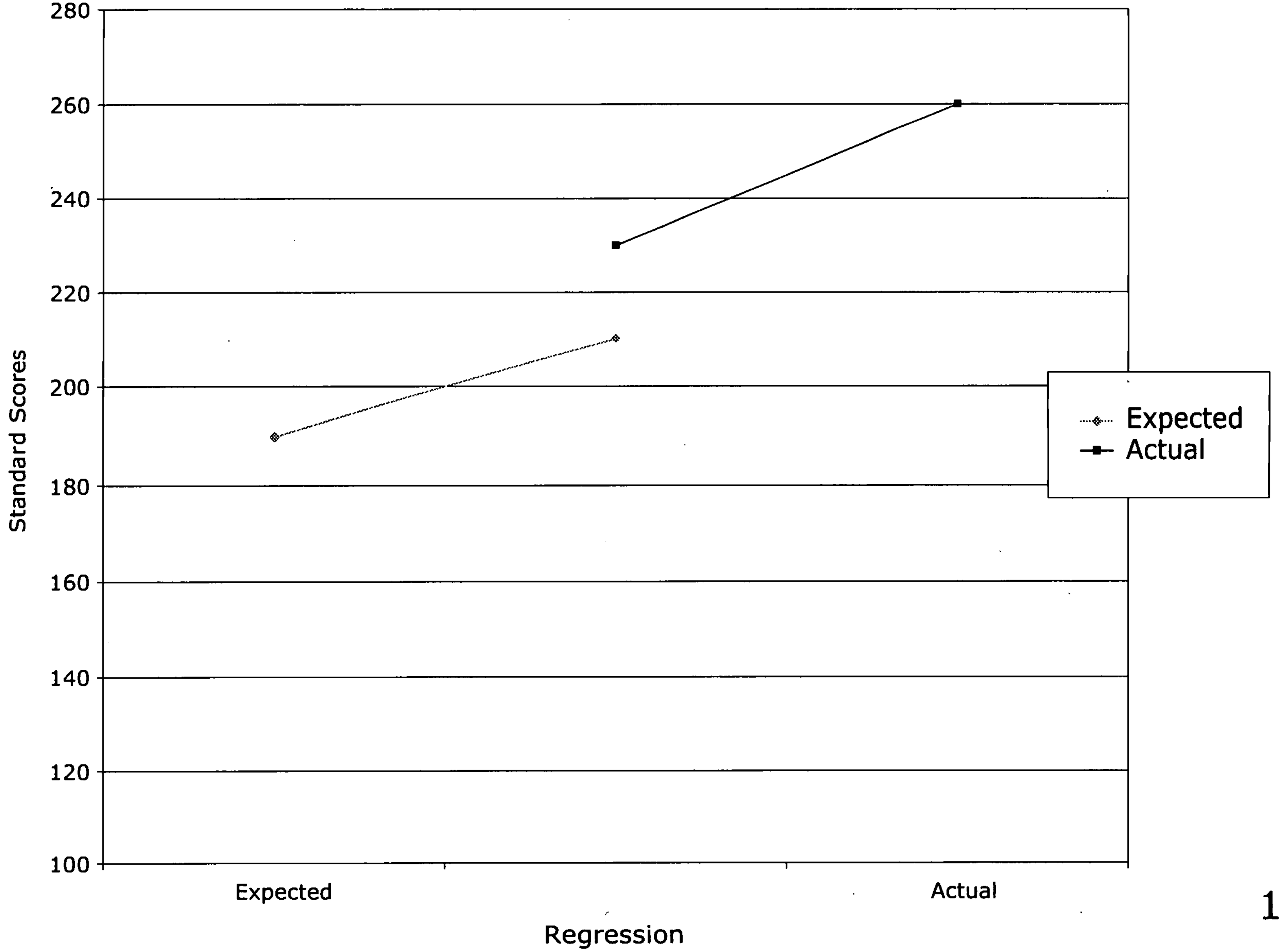
Pretest-Posttest Model



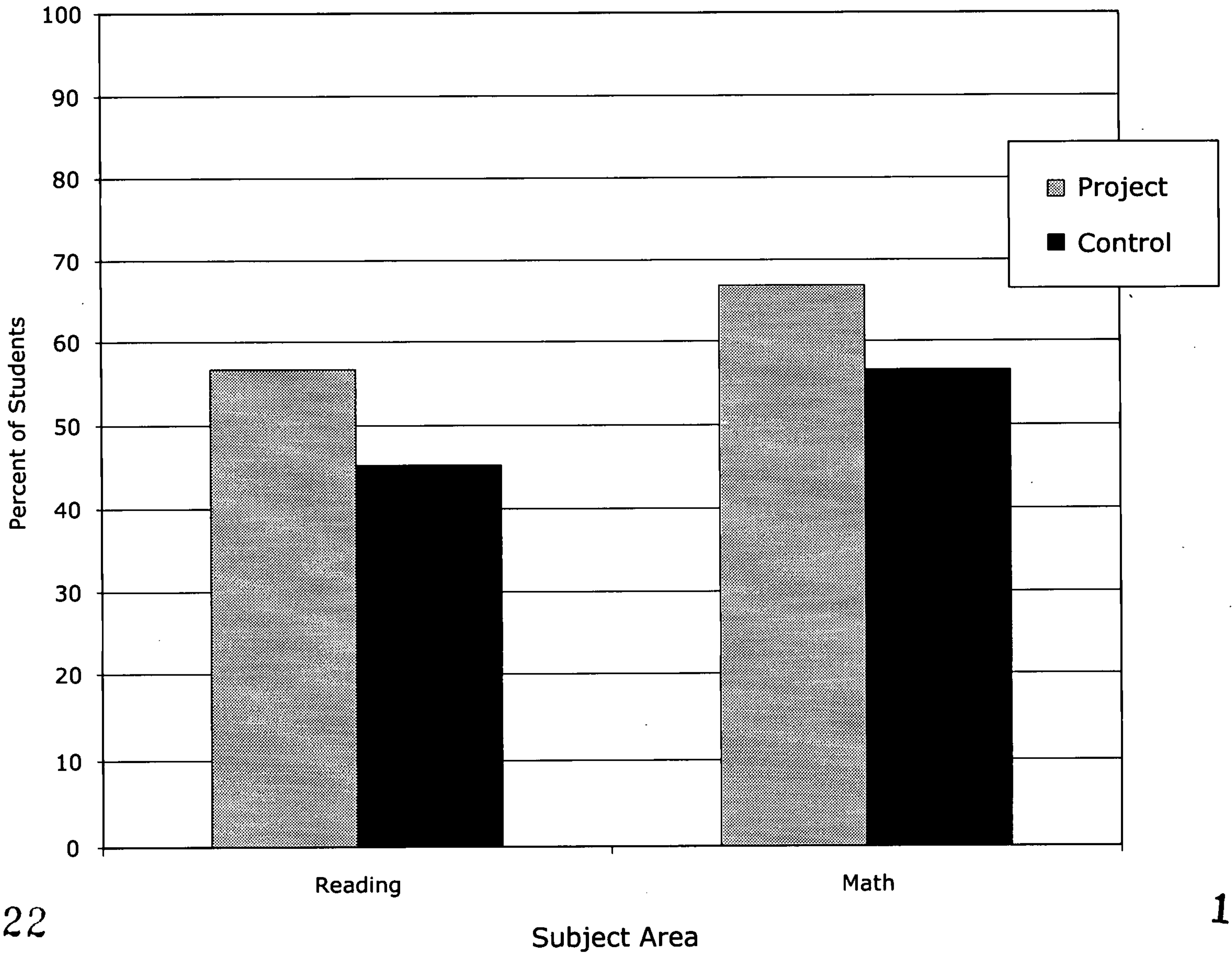
Comparison Group Model



Regression Model



Control Group Model



The Evaluation Process

- What questions do we want to address?
- What do we want to look at?
- How do we collect the data?
- How do we analyze the data?
- How do we interpret the data?
- How do we use data to improve the program?
- Are follow-up actions making a difference?

Data Collection Methods

- Document review
- Questionnaire survey
- Interview
- Focus group
- Observation
- Assessment of student achievement

Data Collection Considerations

- Model selection
- Instrument selection
- Data sources
- Sampling
- Multiple measures
- Data quality

Data Analysis

- Descriptive statistics
- Inferential statistics
- Patterns and trends
- Data disaggregation

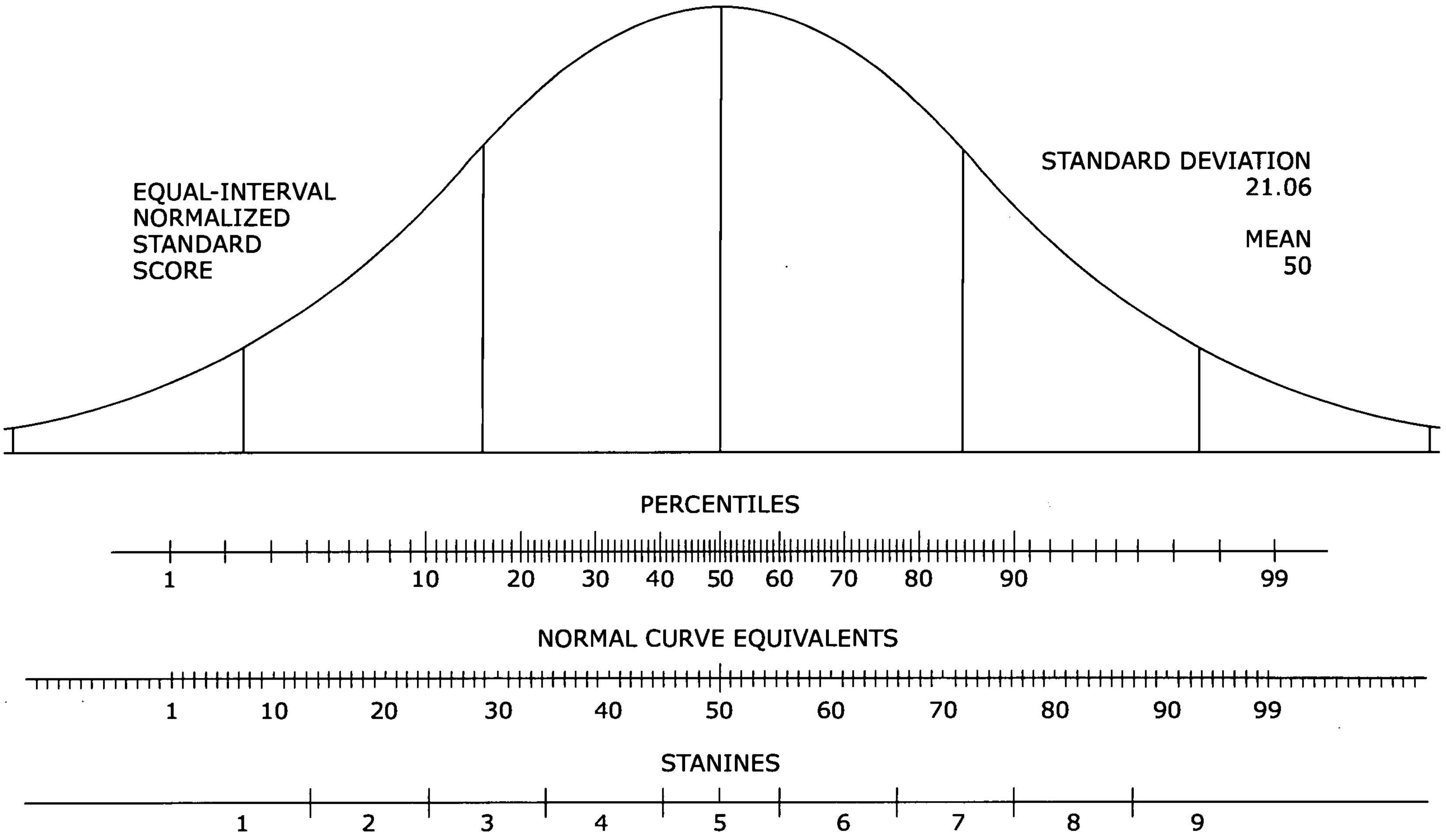
Types of Statistics

- Frequency count
- Percentage
- Mode
- Median
- Mean/standard deviation

Types of Test Scores

- Raw scores
- Percent correct
- Ratings
- Percentiles
- Quartiles
- Stanines
- Normal curve equivalents (NCEs)
- Grade equivalents
- Standard scores

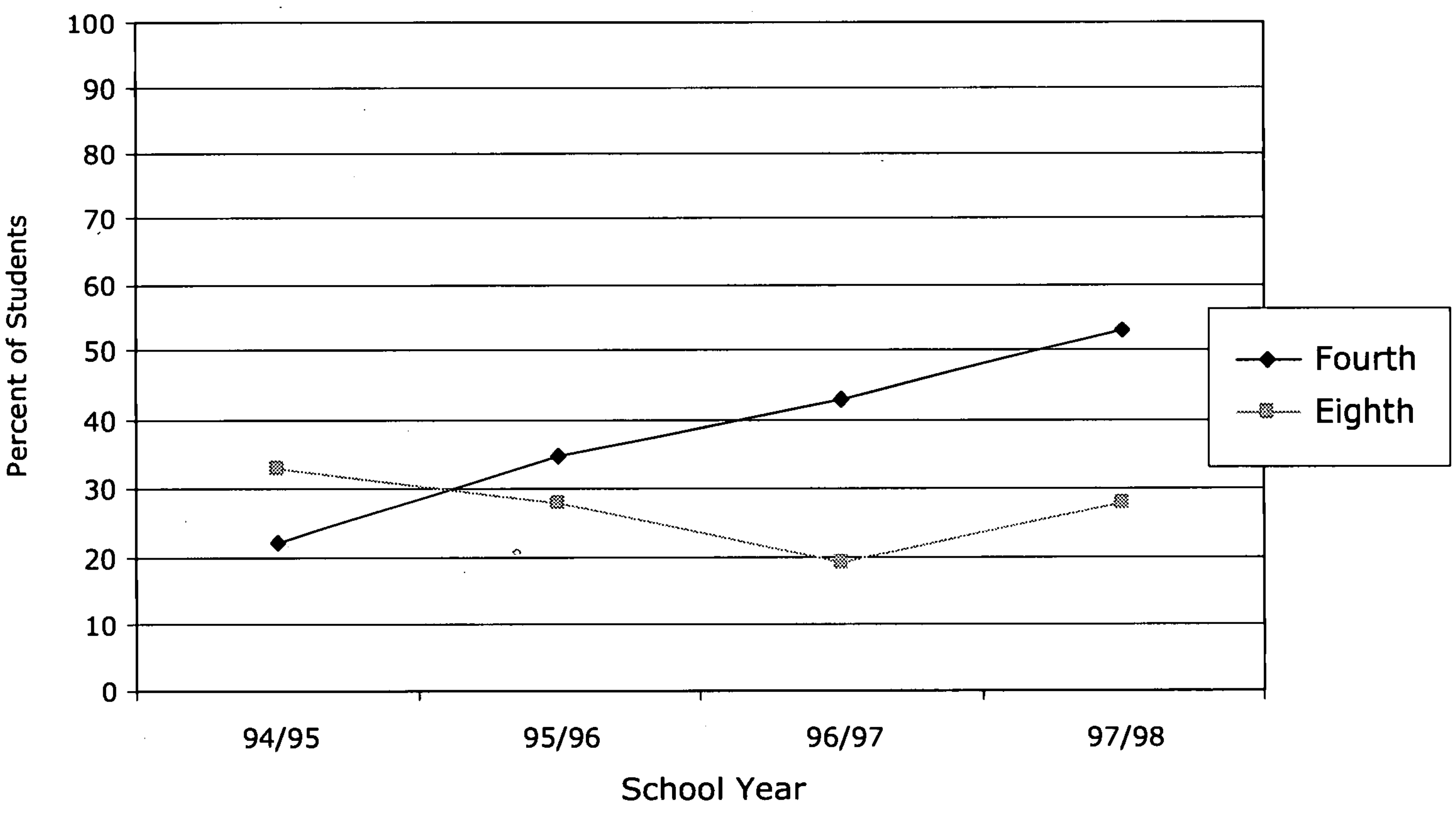
Normal Curve Equivalents



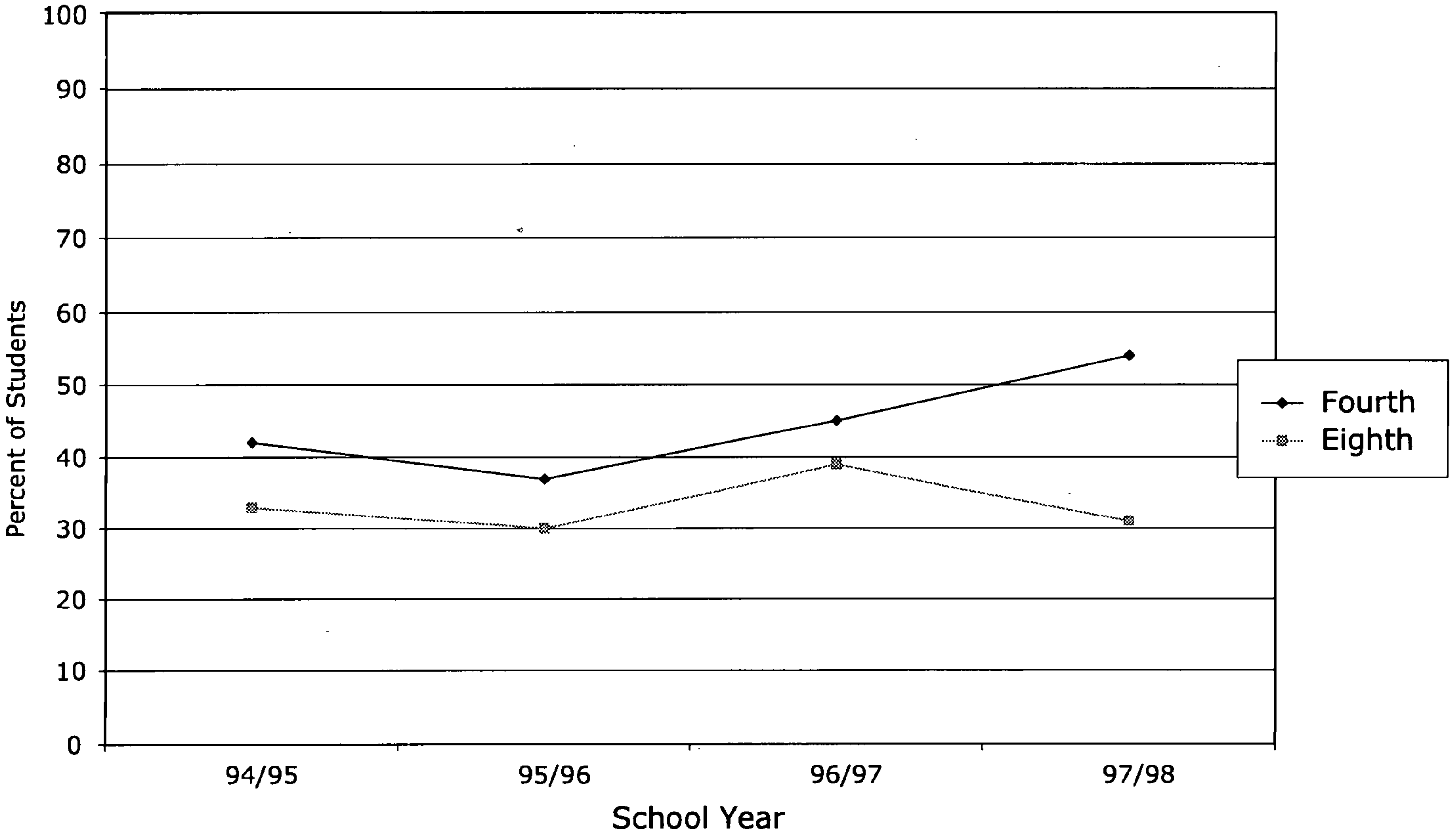
136

137

Percent of Students Meeting Mathematics Benchmarks



Percent of Students Meeting Reading Benchmarks



140

141

Handout: Advantages and Disadvantages of Evaluation Models

Model	Description	Advantages	Disadvantages
Pretest-Posttest	This model provides an expectation of program outcomes based on the current status.	■ Highly feasible in a school setting ■ Shows growth against baseline ■ Shows patterns and trends if conducted longitudinally ■ Can assess relative or absolute growth	■ May lack rigor—difficult to attribute effects to program ■ Difficult to control extraneous factors
Comparison Group	This model provides an expectation of program outcomes based on a comparable group.	■ Relatively strong scientific rigor ■ Can attribute effects to program ■ Can compare progress toward meeting common criteria (e.g., state standards)	■ May be difficult to find a comparable group ■ Selected groups may differ in some important but unknown ways ■ Increased data collection burden
Regression	This model uses a statistical method to predict or project program outcomes	■ Relatively strong scientific rigor ■ Can statistically control for extraneous factors affecting outcomes ■ Does not require existing control or comparison groups	■ Feasibility depends on availability of sufficient archival data ■ Model can be misused ■ Statistical expertise generally not available among existing school/district staff
Control Group	This model provides an expectation of program outcomes based on what happens in an equivalent or control group.	■ Has the strongest scientific rigor with random assignment of students to intervention ■ Can statistically control for extraneous factors affecting outcomes ■ Can attribute effects to program ■ Can compare progress toward meeting common criteria (e.g., state standards)	■ May be difficult, if not impossible, to find an equivalent group ■ Random assignment is typically not feasible in a school setting ■ Increased data collection burden

Handout: Data Collection Matrix

Method	Focus	Advantages	Disadvantages
Document Review	■ Nature and level of school reform activities ■ Incidence of events of interest ■ Existing student achievement information	■ Data already exist ■ Low cost ■ Typically unobtrusive ■ Relatively unbiased	■ Lack of quality control ■ Validity and reliability may be unknown ■ Can be limited in scope
Interview	■ School staff/parent/student perceptions ■ School staff/parent satisfaction ■ Improvement suggestions ■ Degree of implementation ■ Anticipated and unanticipated outcomes	■ Indepth information ■ Quality control ■ High response rate ■ Opportunity to probe	■ Relatively costly ■ Needs trained data collectors ■ Data can be biased ■ May require careful sampling
Survey	■ School staff/parent/student perceptions ■ School staff/parent/student satisfaction ■ Improvement suggestions ■ Degree of implementation ■ Anticipated and unanticipated outcomes	■ Relatively low cost ■ Can include structured and open-ended information ■ Relative ease of administration ■ Can cover a large number of respondents	■ Response rate often a problem ■ Needs careful sampling ■ Data can be biased ■ Open-ended data may be difficult to analyze
Focus Group	■ School staff/parent/student perceptions ■ School staff/parent/student satisfaction ■ Implementation issues ■ Improvement suggestions ■ Degree of implementation ■ Anticipated and unanticipated outcomes	■ Indepth information on program implementation and outcomes ■ Relatively free of response rate problems ■ Interactive discussion among stakeholders	■ Relatively high cost ■ Needs trained facilitators ■ May be difficult to achieve appropriate representation in recruitment of participants ■ Group dynamics can bias discussion
Observation	■ Program implementation ■ Classroom activities ■ Instructional practices ■ School climate	■ Increased objectivity and authenticity of data ■ Can provide contextual information	■ Needs trained observers ■ Relatively high cost ■ Can be obtrusive ■ Often just a snapshot of program implementation ■ May not reflect typical reality
Assessment	■ Student performance in cognitive and affective domains	■ Objective data often with known reliability and validity ■ Can be low cost (standardized testing) ■ Can include large samples of students	■ Provides a generally accepted portrayal of schooling outcomes ■ May provide a limited and narrow picture of student performance ■ Can be high cost (performance-based assessments) ■ May need careful sampling

Small Group Activity #1—Collecting Data

How do we collect data?

Evaluation Question	Data Collection Method	Data Source	Instrument	Date
In what ways is the school/district administration providing support for the school reform effort?				

Activity #2—Analyzing Data

How do we analyze data?

Evaluation Question	Type of Data	Data Analysis Method	Criteria
In what ways are teachers changing and improving their instructional practice?			

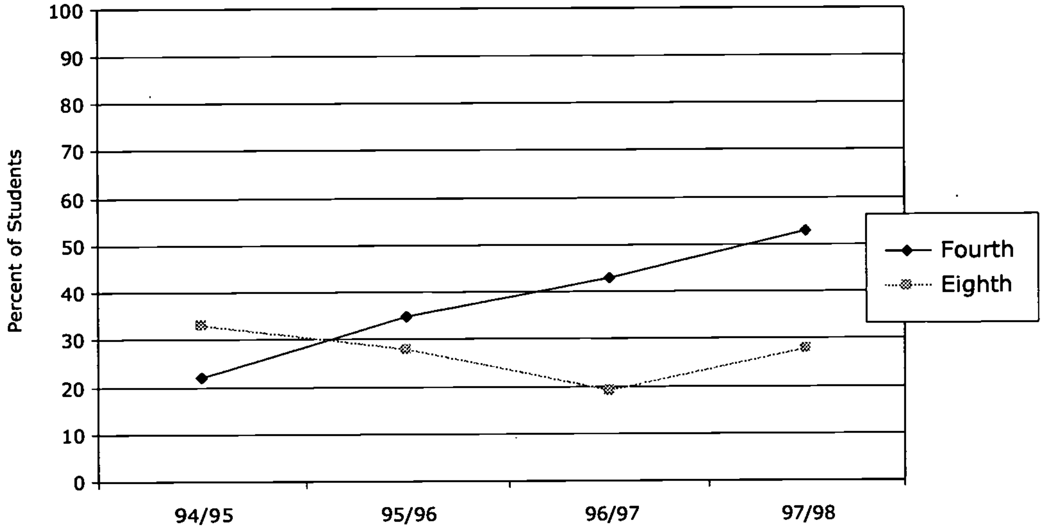
Activity #3—Interpreting Data

What are the data telling us?

Percent of Students Meeting State Benchmarks

Grade (Subject)	94/95	95/96	96/97	97/98
Fourth (Reading)	43	38	46	55
Eighth (Reading)	34	31	40	32
Fourth (Math)	24	36	44	55
Eighth (Math)	35	29	20	29

Percent of Students Meeting Mathematical Benchmarks



Narratives:

1.

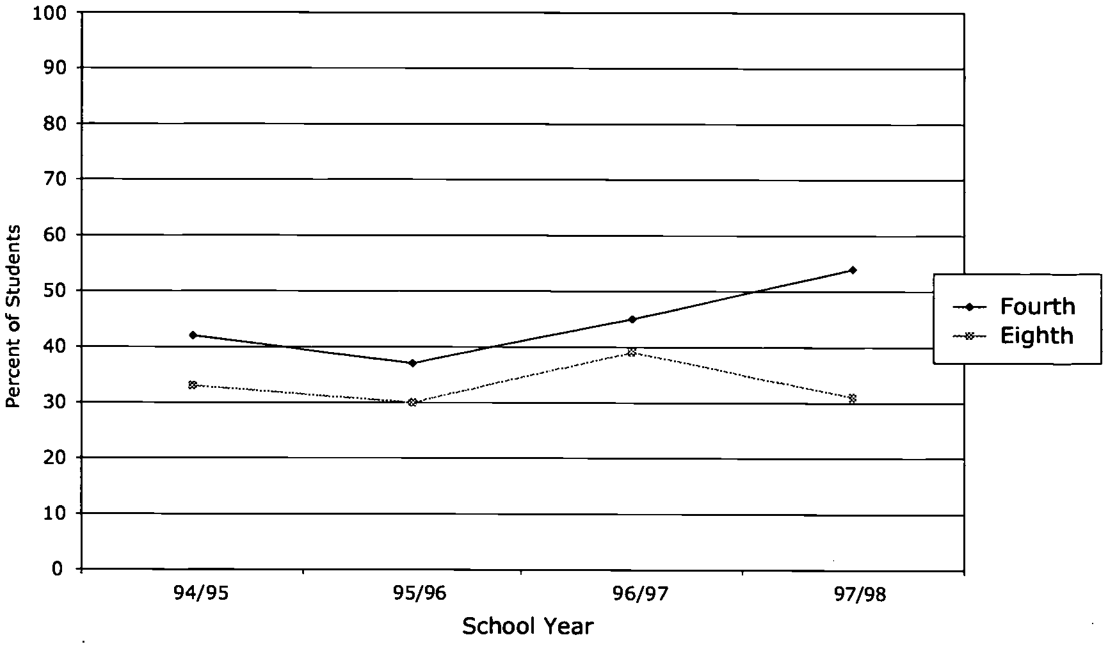
2.

3.

146

Activity #4—Planning Follow-Up

Percent of Students Meeting Reading Benchmarks



147

Activity #4—Planning Follow-Up

Evaluation Question	Key Findings	Action To Be Taken	Person Responsible	Date
Is student performance improving over time?				

148

Design Sample

Instructions to the Presenter

Evaluation of schoolwide projects is needed in order to assess the level and degree of student achievement attributable to change efforts. Various evaluation models, theories, and approaches have proliferated. A single, one-size-fits-all approach to evaluation is difficult, if not impossible to define. Rather, a multiple-method approach will be needed and the methods used will vary from school to school as well. Evaluation is not a single method, design, or approach but a variety of activities from which to pick and choose as appropriate to meet accountability requirements and information needs with available resources. A comprehensive evaluation will provide answers to all parts of the question, "Who does what to whom, with what results, at what costs?" A rigorous evaluation to completely answer this question is typically beyond the resources of most local projects. It is necessary to decide which parts of this question are most relevant and feasible to answer in the schoolwide evaluation effort.

The following activity is designed to help you use the information presented in this guidebook to identify some conceptual distinctions relevant to evaluating schoolwide projects. The type of schoolwide evaluation conducted can range from a simple impact study with little attention paid to implementation issues and a focus on a single measure of student achievement to a complex, fully-designed formative and summative evaluation.

In addition to discussing the strengths and weaknesses of the three evaluation designs, the information provided in this guidebook can be used to determine whether the schools have built a rational cause and effect relationship between the schoolwide model activities and their impact on student achievement. That is, can the school demonstrate that the schoolwide model being implemented has direct relationships to changes in student learning?

Small-Group Activity #5 (40 minutes)

This activity should be completed at the end of the training. Below are three examples of evaluations used by schools interested in identifying the success of their schoolwide project. Break into three small groups, each group taking one of the school scenarios, and discuss the nature of the evaluation using the information provided in this guidebook to answer the questions following each scenario. At the end of the activity, please report the results of your discussion to the full group.

School 1 Scenario

An elementary school with grades kindergarten through sixth implemented a schoolwide reading program—School Improvement Model A—this past year as part of the state's comprehensive school reform initiative. The schoolwide reading model was selected because the school's expected ultimate outcome of children meeting the state reading standards was successfully met in a neighboring school that had implemented the same reading model. Overall, the principal felt the reading scores at his school were dismal; state assessments on writing and math were below the 50th percentile, as well, but the principal thought changes to the entire school curriculum would be too overwhelming for his school staff to endorse.

Support from the schoolwide reading model developers consisted of a week-long training session for 12 of the 15 teachers two weeks before the beginning of school. The focus of the training was how to implement the reading model. Part of the training stressed the importance of completing a checklist of implementation indicators every eight weeks so staff could self-assess how well they were implementing the model's reading components; no other support was provided by the reading model developers. The three teachers who did not receive the staff development training received literature on the newly implemented model and were briefed by those who attended the training. None of the teachers reviewed the grant proposal that was awarded federal funds to implement the school reform model. Additionally, the lone support from the local school district came in the form of funds to implement the specific schoolwide model.

The school evaluation plan took a minimalist approach to identifying model success; increase in student achievement was the sole impact criterion of the school. Baseline data on children's reading scores were at or below the 30th percentile as measured by the California Achievement Test (CAT). The goal of the school was to get 90 percent of the underachieving children to make one and a half years of progress on the reading section of the CAT.

Discussion Questions for Activity #5, School 1

1. What are the strengths of this evaluation?

2. What are the limitations of this evaluation?

3. What would improve the evaluation, at both the formative and summative stages?

4. Will there be evidence for fidelity of model implementation?

5. Is there sufficient evidence collected to demonstrate the school's progress toward its goal?

6. What evaluation model (for example, growth or pretest-posttest) is being utilized? What are the strengths and disadvantages of using this evaluation model?

School 2 Scenario

Staff at School 2 spent one year reviewing their school's strategic plans, the districtwide needs assessment, recent standardized tests, and parent surveys to help identify goals for the coming year in their elementary school. These data helped the school staff decide to implement a schoolwide model to help students become proficient in reading. Along with community members, the school staff felt that implementation of a more structured reading program would prepare students to meet reading standards set by the state and school district.

The school decided it would need to implement a model that would achieve its goals of (1) getting all parents and children involved in the school program, and (2) bringing all students within one grade level in reading as measured by the state standardized test and with 80 percent of the children passing the state benchmark assessment. Based on their desired outcomes, School 2 selected School Improvement Model B to provide the best opportunity for the growth of their students. The staff also felt that the model supplemented its current math and writing curricula. The school also receives financial support and technical assistance from its local school district. The support offers teachers a chance to receive professional development and to obtain the appropriate materials and equipment.

Although the model chosen by the school supported the nine required components of CSRD, little evaluation consideration was given to each of the components. For example, no data are to be collected on sustained support within the school after the initial implementation of the model. However, the staff plan to work with the model developers on data collection surrounding the formative evaluation. Model B contains a schoolwide plan for instruction, assessment, classroom management, professional development, and parent involvement. The model focuses on shared reading, vocabulary building, and writing activities. Teachers have a detailed guide for teaching each component. The staff receive year-round professional development from the model developers. In addition to receiving an initial professional development at the beginning of the school year by the model developers, school component meetings are conducted throughout the year. During the first year of operation the school will receive two implementation checks from the model developers, with two implementation checks conducted during the second year. The model developers will use their own checklists to ensure proper model implementation. Annual curriculum refresher courses are offered to new teachers and anyone else on staff who feels the need for additional training.

The model has specific benchmarks that align well with the state benchmarks. Therefore, the students will be assessed every two months on the model's curriculum-based measure, and those children who show the greatest need will get additional help with their reading. The children are also assessed annually on the school district benchmark, as well as at third and sixth grades on the state benchmark assessment.

Reports are provided to the school staff by the model developers regarding what is going well in the school and next steps that should occur for proper implementation to occur. Data from the state reading test will provide the school staff with indicators of student achievement gains.

At the end of the second year the school will hire a school district evaluator to help them compile, analyze, and interpret the comprehensive implementation data and the district and state benchmark assessments. These data will provide the staff with the information to determine changes in student achievement.

Once the data have been analyzed and interpreted, a report will be provided to the school to make any programmatic changes necessary to further improve students' academic success and improve parent involvement in the school.

Discussion Questions for Activity #5, School 2

1. What are the strengths of this evaluation?

2. What are the limitations of this evaluation?

3. What would improve the evaluation, at both the formative and summative stages?

4. Will there be evidence for fidelity of model implementation?

5. Is there sufficient evidence collected to demonstrate the school's progress toward its goal?

6. What evaluation model (for example, growth or pretest-posttest) is being utilized? What are the strengths and disadvantages of using this evaluation model?

School 3 Scenario

Upon hearing that the state of Oregon would fund 20 Comprehensive School Reform Demonstration (CSRD) sites in the coming year, staff at School 3 began to review their school's strategic plans, the districtwide needs assessment, recent standardized tests, and parent surveys to identify areas in which they could help children perform better in school. These data helped the school staff decide that a new schoolwide model could indeed help their students become more proficient in reading, an area where the latest district assessments indicated School 3's children were performing miserably. Along with community members, the school staff felt that implementing a more structured reading program would prepare students to meet reading standards set by the state and school district. The school staff recently implemented a new schoolwide math model and a new literacy model, and the staff thought the implementation of a new reading model would provide students with the richest of environments in which to learn. After support among school staff was obtained for implementing a new model, a committee of teachers, the principal, and school district staff wrote a proposal for CSRD funding. School staff interested in reviewing the grant were encouraged to offer feedback. Once the proposal was funded, all teachers were required to read the proposal.

The primary goal—as determined by the CSRD Advisory Committee made up of school staff, district staff, and parents of children attending School 3—was for students to become more proficient in reading. Breaking this goal down even further, the measurable objectives were to increase the number of children reading at grade level by 2 percent each year and increase the number of children meeting the Oregon state standard for reading by 10 percent each year. The local school district provided a third-party evaluator to assist in defining measurable goals and to help the staff identify how these goals could be achieved through a schoolwide model. The evaluator assisted in helping the school identify a research-based model that included classroom activities, curriculum, resources, and assessments that would help children perform better in School 3. The model chosen to support children's learning was School Improvement Model C.

Model C contains a schoolwide plan for instruction, assessment, classroom management, professional development, and parent involvement. The model focuses on shared reading, vocabulary building, and writing activities. Teachers have a detailed guide for teaching each component. The staff receives year-round professional development from the model developers. In addition to receiving initial professional development at the beginning of the school year by the model developers, school component meetings are conducted throughout the year. Annual curriculum refresher courses are offered to new teachers and anyone else on staff who feels the need for additional training.

The advisory committee will oversee both the formative and summative evaluation. The committee will meet at least every two months to review the ongoing data collection. During the first year of operation, the school will receive three implementation checks from the model developers, with two implementation checks conducted during the second year. This advisory committee, with the help of the model developers, will create a calendar and checklist to aid in the tracking of appropriate model implementation. Interviews and surveys of students, teachers, and parents will be used to collect information on various aspects of model implementation. Additionally, classroom observations and focus groups with teachers will provide valuable data on how the comprehensive program is being implemented. The advisory committee's goal will be to verify the success of the model implementation and make any modifications to classroom instruction, parent involvement, or other program components.

School 3's evaluation plan will identify progress toward its goal using both state and local data assessments. To measure progress using state assessments, School 3 will use Title I Adequate Yearly Progress Criteria as a measure of academic progress. Local student performance measures are important to School 3 as well. The student performance goal is to improve student achievement in reading with the objective of increasing the percentage of students in grades one through six reading at grade level by the end of the first year of implementation by 2 percent. Multiple measures will be used to assess these changes. For example, local pre- and post-reading assessments will be administered as will the CSRD model's 10-week assessment. The final assessment will be a local literacy assessment to be administered at the beginning and end of the school year. To ensure that the program is on the right track, School 3 created interim benchmarks. The objective of the interim benchmark is to increase the number of students reading at grade level by 0.6 percent each trimester. Students will be assessed with the model's 10-week assessment, the local reading assessment, and nightly reading homework records. Where possible, the assessments will be conducted in the spring and fall. For example, fall and spring assessments on oral reading samples will be conducted to identify changes in student reading strategies and understanding of text.

As is evident, School 3's evaluation plan has two purposes: to document project activities and monitor progress toward expected outcomes and to summarize the overall progress of the plan's effectiveness. School 3 is also concerned that each of the nine CSRD components is addressed in the program evaluation. For each of the nine components, specific processes used to review, monitor, and adjust the school program are included as part of the evaluation plan. Some of the evaluation tools will be administered by the local evaluator, while others will be administered by the CSRD's model developer. Still others will be administered by the advisory committee staff. The tables below offer part of the evaluation of the nine CSRD components.

Component 1:

Effective Research-Based Strategies

Goal
- Implement the CSRD plan successfully
- Align classroom practice to Oregon benchmark

Indicator/Strategy
- Implement strategies as intended by model
- Analysis of change in classroom practice

Measurement
- Monitoring
- Teacher reflections on changes in classroom practices

Who
- Advisory committee
- Model developer

When
- 3 visits per year
- Each term

Component 2:

Comprehensive Design

Goal
- Implement, monitor, and refine CSRD plan on ongoing basis

Indicator/Strategy
- Review progress by checking interim student achievement data

Measurement
- Implementation checklist
- Review and evaluate disaggregated data

Who
- Advisory committee

When
- Each term

Component 3:

Professional Development

Goal
- Implement a professional development plan that results in positive change in reading and parent involvement

Indicator/Strategy
- Ensure full participation in activities
- Change in classroom practices

Measurement
- Attendance at each activity
- Classroom observation

Who
- Advisory committee
- Evaluator

When
- Each term
- Ongoing

Component 5:

School Support

Goal
- Implement a professional development plan that results in positive change in reading and parent involvement

Indicator/Strategy
- Advisory committee will communicate and solicit feedback

Measurement
- Polling of staff by secret ballot to identify continued support of the model

Who
- Evaluator

When
- Annually

Component 6:

Parent and Community Involvement

Goal
- Intact and functioning family support team
- Family participation in 20 minutes of reading homework nightly

Indicator/Strategy
- Weekly team meetings, develop support plans for struggling youth
- Homework with parent signoff sheet

Measurement
- Model monitoring process
- Monitor number of returned assignments

Who
- Advisory committee
- Evaluator

When
- Annually
- Each term

Once the data have been collected, the evaluator will work with the advisory committee on ways to analyze the data. Then, working as a group, they will begin to interpret the data. Once the final report has been completed, the evaluator and a member of the advisory committee will present the findings at a community forum. Although programmatic changes were made throughout the project period, the final report will provide additional evidence for possible changes in program practices.

Discussion Questions for Activity #5, School 3

1. What are the strengths of this evaluation?

2. What are the limitations of this evaluation?

3. What would improve the evaluation, at both the formative (implementation) and summative (impact) stages?

4. Will there be evidence for fidelity of model implementation?

5. Is there sufficient evidence collected to demonstrate the school's progress toward its goal?

6. What evaluation model (such as growth or pretest-posttest) is being utilized? What are the strengths and disadvantages of using this evaluation model? In addition to answering the questions after each scenario, discuss whether the school personnel or evaluator would be able to complete the following schoolwide evaluation worksheet. If information is missing for any

component of the worksheet, discuss whether that information may be important to the school and, if so, what changes in the evaluation design would need to occur to provide sufficient evidence for program success.

Evaluation Framework Schoolwide Evaluation

Basic Evaluation Framework: The following is a brief description of the elements needed for a sound school evaluation design. An evaluation design should express student performance goals. Ideally, the goals highlighted in the evaluation design should encompass, but not be limited to, all existing goals identified by your school in your schoolwide plan. Each identified student performance goal has a *specific objective, strategy for attainment, indicators and benchmarks, and measurement method.*

Finally, discuss in your group whether the school in each of the scenarios has built a rational cause-and-effect relationship between the schoolwide model activities and their impact on student achievement. That is, can the school demonstrate that the model being implemented has a direct relationship to changes in student learning? For example, does a school's evaluation model identify how instructional elements (such as project-based activities or curriculum aligned to standards) relate to expected changes in how students learn, feel, and do in school? Furthermore, does the model identify the types of changes in student performance (such as attendance or problem-solving skills) that lead to attaining the desired standard (say, meeting statewide performance standards)?

Student Level Goal Definitions

Student performance goals—*What do we want students to ultimately achieve?*

A general description of student goals.

Objectives—*What do students need to specifically achieve to accomplish goals?*

A specific, measurable description of student performance that identifies a time frame for achieving goals.

Strategies for attainment—*What do schools have to do to help students accomplish goals and objectives?*

A description of the strategies, means, and methods used by schools to accomplish student performance goals.

Local indicators and benchmarks—*What evidence do we need to demonstrate progress toward goals?*

A specific description of the state, local, and interim indicators and benchmarks to be used to measure progress toward student performance goals and objectives.

Measurement methods—How will we gather the evidence needed to demonstrate successful achievement of goals?

A specific description of the instruments or methods to be used to gather evidence of progress toward attainment of student performance goals and objectives.

Source: *Guidelines for preparing a charter school accountability plan, Massachusetts Department of Education*

Examples of Data Collection Techniques Related to Implementation

Questionnaire/Interviews

- ■ Send out a questionnaire following a staff development activity
- ■ Ask teachers to check off units that they have completed
- ■ Talk with students to determine if the materials are being used in the classroom
- ■ Ask teachers for specific examples of how they integrate their curriculum

Discussion Questions for Activity #5

1. What are the strengths of this evaluation?

2. What are the limitations of this evaluation?

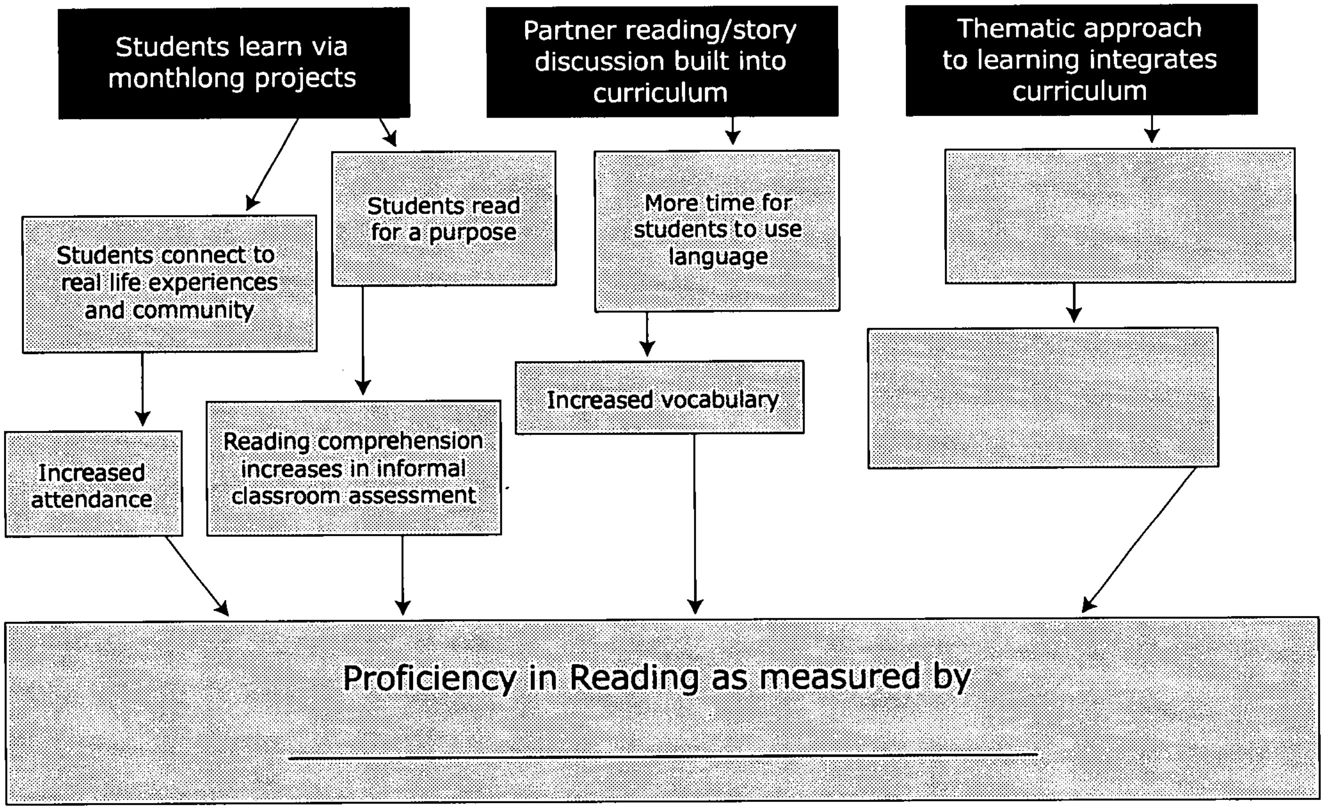
3. What would improve the evaluation, at both the formative (implementation) and summative (impact) stages?

4. Will there be evidence for fidelity of model implementation?

5. Is there sufficient evidence collected to demonstrate the school's progress toward its goal?

6. What evaluation model (growth, pre- post, etc.) is being used? What are the strengths and disadvantages of using this evaluation model?

New Curriculum Implemented


Students learn via monthlong projects
Partner reading/story discussion built into curriculum
Thematic approach to learning integrates curriculum
Students connect to real life experiences and community
Students read for a purpose
More time for students to use language
Increased vocabulary
Increased attendance
Reading comprehension increases in informal classroom assessment
Proficiency in Reading as measured by ____________

Handout: School 1 Scenario

An elementary school with grades kindergarten through sixth implemented a schoolwide reading program—School Improvement Model A—this past year as part of the state's comprehensive school reform initiative. The schoolwide reading model was selected because the school's expected ultimate outcome of children meeting the state reading standards was successfully met in a neighboring school that had implemented the same reading model. Overall, the principal felt the reading scores at his school were dismal; state assessments on writing and math were below the 50th percentile, as well, but the principal thought changes to the entire school curriculum would be too overwhelming for his school staff to endorse.

Support from the schoolwide reading model developers consisted of a week-long training session for 12 of the 15 teachers two weeks before the beginning of school. The focus of the training was how to implement the reading model. Part of the training stressed the importance of completing a checklist of implementation indicators every eight weeks so staff could self-assess how well they were implementing the model's reading components; no other support was provided by the reading model developers. The three teachers who did not receive the staff development training received literature on the newly implemented model and were briefed by those who attended the training. None of the teachers reviewed the grant proposal that was awarded federal funds to implement the school reform model. Additionally, the lone support from the local school district came in the form of funds to implement the specific schoolwide model.

The school evaluation plan took a minimalist approach to identifying model success; increase in student achievement was the sole impact criterion of the school. Baseline data on children's reading scores were at or below the 30th percentile as measured by the California Achievement Test (CAT). The goal of the school was to get 90 percent of the underachieving children to make one and a half years of progress on the reading section of the CAT.

Handout: School 2 Scenario

Staff at School 2 spent one year reviewing their school's strategic plans, the districtwide needs assessment, recent standardized tests, and parent surveys to help identify goals for the upcoming year in their elementary school. These data helped the school staff decide to implement a schoolwide model to help students become proficient in reading. Along with community members, the school staff felt that implementation of a more structured reading program would prepare students to meet reading standards set by the state and school district.

The school decided it would need to implement a model that would achieve its goals of (1) getting all parents and children involved in the school program and, (2) bringing all students within one grade level in reading as measured by the state standardized test and with 80 percent of the children passing the state benchmark assessment. Based on their desired outcomes, School 2 selected School Improvement Model B to provide the best opportunity for the growth of their students. The staff also felt that the model supplemented its current math and writing curricula. The school also receives financial support and technical assistance from its local school district. The support offers teachers a chance to receive professional development and to attain the appropriate materials and equipment.

Although the model chosen by the school supported the nine required components of CSRD, little evaluation consideration was given to each of the components. For example, no data are to be collected on sustained support within the school after the initial implementation of the model. However, the staff plan to work with the model developers on data collection surrounding the formative evaluation. Model B contains a schoolwide plan for instruction, assessment, classroom management, professional development, and parent involvement. The model focuses on shared reading, vocabulary building, and writing activities. Teachers have a detailed guide for teaching each component. The staff receive year-round professional development from the model developers. In addition to receiving an initial professional development at the beginning of the school year by the model developers, school component meetings are conducted throughout the year. During the first year of operation the school will receive two implementation checks from the model developers, with two implementation checks conducted during the second year. The model developers will use their own checklists to ensure proper model implementation. Annual curriculum refresher courses are offered to new teachers and anyone else on staff who feels the need for additional training.

The model has specific benchmarks that align well with the state benchmarks. Therefore, the students will be assessed every two months on the model's curriculum-based measure, and those children who show the greatest need will get additional help with their reading. The children are also assessed annually on the school district benchmark, as well as at third and sixth grades on the state benchmark assessment.

Reports are provided to the school staff by the model developers regarding what is going well in the school and next steps that need to occur for proper implementation to occur. Data from the state reading test will provide the school staff with indicators of student achievement gains.

At the end of the second year the school will hire a school district evaluator to help them compile, analyze, and interpret the comprehensive implementation data and the district and state benchmark assessments. These data will provide the staff with the information to determine changes in student achievement.

Once the data have been analyzed and interpreted, a report will be provided to the school to make any programmatic changes necessary to further improve students' academic success and improve parent involvement in the school.

Handout: School 3 Scenario

Upon hearing that the state of Oregon would fund 20 Comprehensive School Reform Demonstration (CSRD) sites in the coming year, staff at School 3 began to review their school's strategic plans, the districtwide needs assessment, recent standardized tests, and parent surveys to identify areas in which they could help children perform better in school. These data helped the school staff decide that a new schoolwide model could indeed help their students become more proficient in reading, an area where the latest district assessments indicated School 3's children were performing miserably. Along with community members, the school staff felt that implementation of a more structured reading program would prepare students to meet reading standards set by the state and school district. The school staff recently implemented a new schoolwide math model and a new literacy model, and the staff thought the implementation of a new reading model would provide students with the richest of environments in which to learn. After support among school staff was attained for implementing a new model, a committee of teachers, the principal, and school district staff wrote a proposal for CSRD funding. School staff interested in reviewing the grant were encouraged to offer feedback. Once the proposal was funded, all teachers were required to read the proposal.

The primary goal—as determined by the CSRD Advisory Committee made up of school staff, district staff, and parents of children attending School 3—was for students to become more proficient in reading. Breaking this goal down even further, the measurable objectives were to increase the number of children reading at grade level by 2 percent each year and increase the number of children meeting the Oregon state standard for reading by 10 percent each year. The local school district provided a third-party evaluator to assist in defining measurable goals and to help the staff identify how these goals could be achieved through a schoolwide model. The evaluator assisted in helping the school identify a research-based model that included classroom activities, curriculum, resources, and assessments that would help children perform better in School 3. The model chosen to support children's learning was School Improvement Model C.

Model C contains a schoolwide plan for instruction, assessment, classroom management, professional development, and parent involvement. The model focuses on shared reading, vocabulary building, and writing activities. Teachers have a detailed guide for teaching each component. The staff receives year-round professional development from the model developers. In addition to receiving initial professional development at the beginning of the school year by the model developers, school component meetings are conducted throughout the year. Annual curriculum refresher courses are offered to new teachers and anyone else on staff who feels the need for additional training.

The advisory committee will oversee both the formative and summative evaluation. The committee will meet at least every two months to review the ongoing data collection. During the first year of operation, the school will receive three implementation checks from the model developers, with two implementation checks conducted during the second year. This advisory committee, with the help of the model developers, will create a calendar and checklist to aid in the tracking of appropriate model implementation. Interviews and surveys of students, teachers, and parents will be used to collect information on various aspects of model implementation. Additionally, classroom observations and focus groups with teachers will provide valuable data on how the comprehensive program is being implemented. The advisory committee's goal will be to verify the success of the model implementation and make any modifications to classroom instruction, parent involvement, or other program components.

School 3's evaluation plan will identify progress toward its goal using both state and local data assessments. To measure progress using state assessments, School 3 will use Title I Adequate Yearly Progress Criteria as a measure of academic progress. Local student performance measures are important to School 3 as well. The student performance goal is to improve student achievement in reading with the objective of increasing the percentage of students in grades one through six reading at grade level by the end of the first year of implementation by 2 percent. Multiple measures will be used to assess these changes. For example, local pre- and post-reading assessments will be administered as will the CSRD model's 10-week assessment. The final assessment will be a local literacy assessment to be administered at the beginning and end of the school year. To ensure that the program is on the right track, School 3 created interim benchmarks. The objective of the interim benchmark is to increase the number of students reading at grade level by 0.6 percent each trimester. Students will be assessed with the model's 10-week assessment, the local reading assessment, and nightly reading homework records. Where possible, the assessments will be conducted in the spring and fall. For example, fall and spring assessments on oral reading samples will be conducted to identify changes in student reading strategies and understanding of text.

As is evident, School 3's evaluation plan has two purposes: to document project activities and monitor progress toward expected outcomes and to summarize the overall progress of the plan's effectiveness. School 3 is also concerned that each of the nine CSRD components is addressed in the program evaluation. For each of the nine components, specific processes used to review, monitor, and adjust the school program are included as part of the evaluation plan. Some of the evaluation tools will be administered by the local evaluator, while others will be administered by the CSRD's model developer. Still others will be administered by the advisory committee staff. The tables below offer part of the evaluation of the nine CSRD components.

Component 1:

Effective Research-Based Strategies

Goal
- Implement the CSRD plan successfully
- Align classroom practice to Oregon benchmark

Indicator/Strategy
- Implement strategies as intended by model
- Analysis of change in classroom practice

Measurement
- Monitoring
- Teacher reflections on changes in classroom practices

Who
- Advisory committee
- Model developer

When
- 3 visits per year
- Each term

Component 2:

Comprehensive Design

Goal
- Implement, monitor, and refine CSRD plan on ongoing basis

Indicator/Strategy
- Review progress by checking interim student achievement data

Measurement
- Implementation checklist
- Review and evaluate disaggregated data

Who
- Advisory committee

When
- Each term

Component 3:

Professional Development

Goal
- Implement a professional development plan that results in positive change in reading and parent involvement

Indicator/Strategy
- Ensure full participation in activities
- Change in classroom practices

Measurement
- Attendance at each activity
- Classroom observation

Who
- Advisory committee
- Evaluator

When
- Each term
- Ongoing

Handout: School 3 Scenario Continued

Once the data have been collected, the evaluator will work with the advisory committee on ways to analyze the data. Then, working as a group, they will begin to interpret the data. Once the final report has been completed, the evaluator and a member of the advisory committee will present the findings at a community forum. Although programmatic changes were made throughout the project period, the final report will provide additional evidence for possible changes in program practices.

Component 5:

School Support

Goal

■ Implement a professional development plan that results in positive change in reading and parent involvement

Indicator/Strategy

■ Advisory committee will communicate and solicit feedback

Measurement

■ Polling of staff by secret ballot to identify continued support of the model

Who

■ Evaluator

When

■ Annually

Component 6:

Parent and Community Involvement

Goal

■ Intact and functioning family support team

■ Family participation in 20 minutes of reading homework nightly

Indicator/Strategy

■ Weekly team meetings, develop support plans for struggling youth

■ Homework with parent signoff sheet

Measurement

■ Model monitoring process

■ Monitor number of returned assignments

Who

■ Advisory committee

■ Evaluator

When

■ Annually

■ Each term

Handout: Discussion Questions for Activity 5

1. What are the strengths of this evaluation?

2. What are the limitations of this evaluation?

3. What would improve the evaluation, at both the formative (implementation) and summative (impact) stages?

4. Will there be evidence for fidelity of model implementation?

5. Is there sufficient evidence collected to demonstrate the school's progress toward its goal?

6. What evaluation model (such as growth, pretest-posttest) is being utilized? What are the strengths and disadvantages of using this evaluation model?

Handout: Schoolwide Evaluation Worksheet

Basic Evaluation Framework: The following is a brief description of the elements needed for a sound school evaluation design. An evaluation design should express student performance goals. Ideally, the goals highlighted in the evaluation design should encompass, but not be limited to, all existing goals identified by your school in your schoolwide plan. Each identified student performance goal has a *specific objective, strategy for attainment, indicators and benchmarks,* and *measurement method.*

Student Level Goal Definitions

Student performance goals—*What do we want students to ultimately achieve?*

A general description of student goals.

Objectives—*What do students need to specifically achieve to accomplish goals?*

A specific, measurable description of student performance that identifies a time frame for achieving goals.

Strategies for attainment—*What do schools have to do to help students accomplish goals and objectives?*

A description of the strategies, means, and methods used by schools to accomplish student performance goals.

Local indicators and benchmarks—*What evidence do we need to demonstrate progress toward goals?*

A specific description of the state, local, and interim indicators and benchmarks to be used to measure progress toward student performance goals and objectives.

Measurement methods—How will we gather the evidence needed to demonstrate successful achievement of goals?

A specific description of the instruments or methods to be used to gather evidence of progress toward attainment of student performance goals and objectives.

Source: *Guidelines for preparing a charter school accountability plan, Massachusetts Department of Education*

Resources

Below is a listing of useful print and online information resources that relate to evaluating schoolwide reform programs, and a listing of technical assistance providers. Most of the print resources may be borrowed from the Comprehensive Center's Resource Center. Please contact the Comprehensive Center for more information.

Print

Data Use Tools

Bernhardt, V.L. (1998). *Data analysis for comprehensive schoolwide improvement.* Larchmont, NY: Eye on Education.

Targeted at non-statisticians, this practical toolbook shows educators how to gather, analyze, and use data to improve all aspects of schools.

Holcomb, E.L. (1999). *Getting excited about data: How to combine people, passion, and proof.* Newbury Park, CA: Corwin Press.

This practical manual answers questions about what data to collect, how to analyze data, and how to use the data to align school improvement.

Levesque, K., Bradby, D., Rossi, K., & Teitelbaum, P. (1998). *At your fingertips: Using everyday data to improve schools.* Berkeley, CA: MPR Associates, Berkeley, CA: National Center for Research in Vocational Education, & Arlington, VA: American Association of School Administrators.

This workbook is designed to help educators use a variety of data to better manage, monitor, and improve schools. The workbook is structured to help teams and individuals develop performance indicator systems that can be used to identify strengths and weaknesses and to develop strategies to meet educational goals.

Roza, M. (1998). *A toolkit for using data to improve schools: Raise student achievement by incorporating data analysis in school planning.* Newton, MA: Education Development Center, New England Comprehensive Assistance Center.

The *Toolkit* is intended for use by school and district staff interested in using data to improve school programs. This resource will enable users to collect, understand, and use data for creating and improving schoolwide plans designed to increase student achievement. The Toolkit comes with a companion resource, the Data Templates, designed to help collect, disaggregate, and display baseline data.

Wagner, M., Fiester, L., Reisher, E., Murphy, D., & Golan, S. (1997). *Making information work for you: A guide for collecting good information and using it to improve comprehensive strategies for children, families, and communities.* Washington, DC: U.S. Department of Education.

This evaluator's toolkit provides evaluation methods and instruments that schools can use to collect sound information and document program progress. Suggestions are included for starting the evaluation process and documenting results.

Evaluation Tools

Beyer, B.K. (1995). *How to conduct a formative evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development.

This book describes how to conduct an evaluation of educational programs by assessing the program during various stages of its development. The author provides practical checklists, data-collection instruments, and other resources to assist in conducting the evaluation.

Billig, S.H., & Kraft, N.P. (1996). *Linking Title I and service-learning: A planning, implementation, and evaluation guide*. Denver, CO: RMC Research Corporation.

This guide provides guidelines for program planning, operations, and evaluations for Title I programs. Section IV discusses how to evaluate the impact of a program and how to improve its effectiveness.

Cicchinelli, L.F., & Barley, Z. (1999). *Evaluating for success. Comprehensive school reform: An evaluation guide for districts and schools.* Aurora, CO: Mid-continent Research for Education and Learning.

This guide provides practical information, tips, and tools to help schools and districts meet the evaluation requirements of the federally sponsored Comprehensive School Reform Demonstration (CSRD) program. The guide is also useful for schools and districts involved in other comprehensive school reform efforts and especially useful for those who don't have extensive evaluation experience.

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

This comprehensive guide offers educators step-by-step procedures and practical guidance needed to conduct sensible assessments and evaluations, and record and measure progress. It also instructs the reader on how to use evaluation information to aid in school planning and improve management decisions.

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

This is part of the Sage series called The Program Evaluation Kit (2nd ed.). The series contains nine books written to guide and assist practitioners in planning and managing evaluations: (1) *Evaluators handbook;* (2) *How to focus an evaluation;* (3) *How to design a program evaluation;* (4) *How to use qualitative methods in evaluation;* (5) *How to assess program implementation;* (6) *How to measure attitudes;* (7) *How to measure performance and use tests;* (8) *How to analyze data;* and (9) *How to communicate evaluation findings.*

Pechman, E., Allen, S., Funkhouser, J., Kelliher, K., Rouk, U., & Rusnak, K. (1998). *Implementing schoolwide programs: Volume 1, an idea book on planning.* Washington, DC: U.S. Department of Education.

This book focuses on the issues of schoolwide program planning and combining resources. It contains many examples from various schools that illustrate the issues discussed. Thorough assessment of needs and schoolwide planning are essential for comprehensively upgrading the effectiveness of a school. Two appendices provide tools for planning schoolwide programs and extensive information about print, video, and Internet resources available to planners.

RMC Research Corporation. (1995). *Schoolwide programs: A planning manual.* Portland, OR: Author.

Designed to help educators collect data on their school, and plan and implement a schoolwide program. This manual discusses the vision behind and advantages of a schoolwide program. It highlights a four-step process for planning a schoolwide program: (1) conducting a comprehensive needs assessment; (2) managing the inquiry process; (3) designing the schoolwide program; and (4) evaluating the program.

Sanders, J.R. (1992). *Evaluating school programs: An educator's guide.* Newbury Park, CA: Corwin Press.

Here is a general guide to help in planning and conducting school program evaluations. The author guides the reader through each step in the evaluation process: how to focus the evaluation, and how to collect, organize, analyze, report, and use the information collected.

Examples of State CSR Evaluation Plans

Oregon Department of Education. (1999). *Oregon Comprehensive School Reform Demonstration Program 1999 state evaluation plan: Guidance and timeline.* Salem, OR. Author

The plan has two purposes: to document project activities and progress toward expected outcomes, and to summarize the overall progress of the reform program.

Washington State Office of Superintendent of Public Instruction. (1999). *Comprehensive School Reform Demonstration Program local evaluation report.* Olympia, WA. Author.

The purpose of this evaluation report is to "monitor and document CSRD program implementation; to assess progress toward expected outcomes; and to determine overall program effectiveness in improving student achievement."

For information about other state CSR evaluation plans, contact the state departments of education.

Research Articles and Studies

Glennan, T.K., Jr. (1998). *New American Schools after six years.* Santa Monica, CA: RAND.

In July of 1991, New American Schools (NAS) was established to develop designs for what were termed "break the mold" schools. Its initial goal was to create designs to help schools enable students to reach high educational standards. It then moved to implement the new design in a significant number of schools as an element of a strategy for promoting wider education reform. This report describes RAND's perspectives on the evolution of NAS' mission.

Kushman, J.W., & Yap, K.O. (1999). What makes the difference in school improvement? An impact study of Onward to Excellence in Mississippi schools. *Journal of Education for Students Placed at Risk, 4*(3), 277-298.

The study examined the implementation of OTE and its impact on student achievement over a five-year period. The study concludes that implementation and retention were uneven across schools and that high-fidelity implementation appears to lead to positive results. The authors discuss the difficulties in implementing whole-school reform models and the factors that help or hinder success.

Stringfield, S., Datnow, A., Ross, S.M., & Snively, F. (1998). Scaling up school restructuring in multicultural, multilingual contexts: Early observations from Sunland County. *Education and Urban Society, 30*(3), 326-357.

This study addresses three policy questions: (1) How effective are current school restructuring programs in improving the achievement of students in schools with large numbers of language-minority students? (2) Are some models better suited to multilingual environments than others? (3) What actions at the federal, state, district, and school level increase or decrease the probability of these schools obtaining full benefits from these models?

Taylor, D.L., & Teddlie, C. (1999). Implementation fidelity in Title I schoolwide programs. *Journal of Education for Students Placed At Risk, 4*(3), 299-319.

This study examines the extent to which schools that received Title I funds for schoolwide programs implemented the plans they developed. Findings showed that while schools implemented some of the plan components, such as hiring Title I teachers and teaching assistants, instructional innovations included in the plans were not implemented. The article concludes with specific recommendations for districts and schools.

Wong, K.K., & Meyer, S.J. (1998). Title I schoolwide programs: A synthesis of findings from recent evaluation. *Educational Evaluation and Policy Analysis, 20*(2), 115-136.

This article synthesizes what is known about Title I schoolwide programs, focusing on programmatic and organizational characteristics of schoolwide program schools and districts, and evidence of the effectiveness of schoolwide program schools, especially in terms of student performance.

Online Publications and Resources

Herman, R., Aladjem, D., McMahon, P., Masem, E., Mulligan, I., O'Malley, A.S., Quinones, S., Reeve, A., & Woodruff, D. (1999). *An educator's guide to schoolwide reform.* Arlington, VA: Educational Research Service. Retrieved June 14, 2000 from the World Wide Web: www.aasa.org/Reform/index.htm

The American Institutes for Research (AIR) developed this guide for educators and others to use when investigating different approaches to school reform. It reviews the research on 24 "whole-school," "comprehensive," or "schoolwide" approaches.

Comprehensive School Reform Demonstration [Web site] Northwest Regional Educational Laboratory, Portland, OR www.nwrel.org/csrdp/index.html

This Web site offers descriptions of school reform models, contact information for service providers, a listing of Northwest school CSR sites, descriptions of the types of assistance available, and Internet links to articles about reform models.

Comprehensive School Reform Demonstration Program [Web site] U.S. Department of Education, Washington, DC www.ed.gov/offices/OESE/compreform/

This Web site includes a publications list, tools, state contacts, and other Web-site links related to CSRD.

Klein, S., Medrich, E., & Perez-Ferreiro, V. (1996). *Fitting the pieces: Education reform that works.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement. Retrieved June 14, 2000 from the World Wide Web: www.ed.gov/pubs/SER/FTP

An indepth study of 12 education reform studies commissioned by the U.S. Department of Education. Each study comprises three volumes. Volume I contains a discussion of the study, case study summaries of the schools or school districts examined, and recommendations. Volume II contains detailed case studies. Volume III is a technical appendix explaining the study's methodology.

NWREL's Assessment & Evaluation Services [Web site]. Northwest Regional Educational Laboratory, Portland, OR www.nwrel.org/eval/index.html

The Assessment and Evaluation Program translates for educators and community leaders the best research into practical, user-friendly resources and services for the assessment of educational results. The Web site contains a searchable database of assessment resources available for loan through the Assessment Resource Library.

Quellmalz, E., Shields, P.M., Knapp, M.S., Bamburg, J.D., Anderson, L., Hawkins, E., Hill, L., Ruskus, J., & Wilson, C.L. (1995). *School-based reform: Lessons from a national study. A guide for school reform teams.* Washington, DC: U.S. Department of Education. Retrieved June 14, 2000 from the World Wide Web: http://ed.gov/pubs/Reform/index.html

This national study, conducted by SRI International for the Planning and Evaluation Service of the U.S. Department of Education, examined effective school programs and other school-based reform efforts nationwide. This guide provides advice and specific examples based on the findings of the study.

Videotape

Ross, S., & Davis, D. (Presenters). (1999). *Selecting and implementing comprehensive school reform programs* [Videotape]. Portland, OR: Northwest Regional Educational Laboratory, Comprehensive School Reform Demonstration Program.

This videotape provides detailed information on keys to selecting, implementing, and evaluating Comprehensive School Reform Programs. Dr. Stephen Ross of the University of Memphis discusses the formative evaluation process for school reform programs and presents examples of evaluation instruments.

Technical Assistance Providers

U.S. Department of Education Regional Offices

The U.S. Department of Education maintains 10 regional offices throughout the country. The following offices have representatives in each regional office:

The Secretary's Regional Representative (SRR) and staff conduct departmental business on many issues. The Office of Postsecondary Education (OPE) handles questions related to student financial assistance programs. The Office of Special Education and Rehabilitative Services (OSERS) assists constituents with rehabilitative services. The Office for Civil Rights (OCR) responds to questions about, and reviews complaints related to, civil rights issues. The Office of the Inspector General (OIG) investigates potential violations of law and conducts audits on Department-funded programs. The Office of Management (OM) has personnel offices or representatives in each of the regional offices.

Additional information regarding the Regional Offices can be found at: www.ed.gov/pubs/TeachersGuide/offices.html or by contacting the U.S. Department of Education.

Comprehensive Regional Assistance Centers (CCs)

The 15 Comprehensive Centers provide comprehensive training, technical assistance, and capacity building to local education agencies, schools, tribes, states, and community-based organizations. Services are designed to help schools and districts focus on improving teaching and learning, especially in the development of schoolwide programs and programs that improve the opportunity for all children to meet challenging state content and student performance standards. These services include meeting the special needs of children served under the Improving America's Schools Act (IASA), including children in high-poverty schools, migrant children, immigrant children, Native American children, children with limited English proficiency, neglected or delinquent children, homeless children, and children with disabilities.

Additional information regarding the Comprehensive Assistance Centers can be found at: www.wested.org/cc/html/ccnetwork.htm or by contacting the U.S. Department of Education.

Regional Educational Laboratories

The Regional Educational Laboratory Program is the U.S. Department of Education's largest research and development investment, designed to help educators, policymakers, and communities improve schools and help all students attain their full potential. The network of 10 Laboratories works to ensure that those involved in educational improvement at the local, state, and regional levels have access to the best available research and knowledge from practice. A main priority that guides all Laboratory work is helping educators and administrators expand systemic reform to benefit schools, and the educational programs within them, in all communities.

Additional information regarding the Regional Educational Laboratories can be found at: www.relnetwork.org or by contacting the U.S. Department of Education.

Eisenhower Regional Math/Science Consortia

The 10 consortia provide technical assistance and disseminate information to teachers and other educators in implementing mathematics and science programs in accordance with state standards.

For information on service providers in your region, please contact your state department of education or the U.S. Department of Education.

Additional information regarding the consortia can be found at www.enc.org or by contacting the U.S. Department of Education.