Evaluating ARRA Programs and Other Educational Reforms: A Guide for States

April 2011

Prepared By:

Irma Perez-Johnson, Mathematica Policy Research, Inc.
Kirk Walters, American Institutes for Research
Michael Puma, Chesapeake Research Associates

And

Rebecca Herman
Michael Garet
Jessica Heppen
Mariann Lemke
American Institutes for Research
Daniel Aladjem
SRI International

Samia Amin
John Burghardt
Mathematica Policy Research, Inc.

The authors gratefully acknowledge funding from the Institute of Education Sciences (IES) of the U.S. Department of Education for development of this guide. The views expressed in this document are those of the authors and do not represent those of IES or the U.S. Department of Education.

Suggested Citation:

Perez-Johnson, Irma, Kirk Walters, Michael Puma and others. "Evaluating ARRA Programs and Other Educational Reforms: A Guide for States." Resource document developed jointly by The American Institutes for Research and Mathematica Policy Research, Inc. April 2011.

CONTENTS

[This page was intentionally left blank for double-sided printing.]

Introduction

byKirk Walters, Irma Perez-Johnson, Michael Puma and Mariann Lemke

The American Institutes for Research (AIR) and Mathematica Policy Research (MPR) developed this guide to help you consider evaluation issues likely to arise as you launch ARRA-funded initiatives and other educational reform activities. Many states are already involved in evaluation, so many of the ideas presented here may be familiar. We hope that this guide will provide additional information and straightforward strategies to help you integrate evaluation into your educational reform efforts.

The American Recovery and Reinvestment Act (ARRA) is providing states like yours and districts important opportunities to strengthen efforts to improve educational outcomes for all children. ARRA can support efforts to implement new comprehensive reform initiatives or to enhance or expand existing innovative programs. Four specific reform areas are emphasized:

- Revamping teacher evaluation systems to include measures of student learning
- Ensuring that school leaders and teachers have the professional development and support opportunities they need to be successful
- Creating incentives to place effective teachers in high-need schools and subject areas
- Developing comprehensive support systems to help turn around low-performing schools

Although your primary focus will be on making these reform efforts a success, what you learn from these experiences will help guide future policy. Structuring and implementing ARRA-funded initiatives and other education reforms in evaluation-friendly ways will help you understand the results of your efforts and provide the information you will need for future policy and program decisions. Valuable learning opportunities will be lost if you do not ask these questions and analyze data about your reform programs. Important questions to consider include:

- After the funds are spent, will you know exactly what was done, how wide the scope of participation was, what worked well, and what unexpected implementation problems were encountered?
- How will you know whether your programs and initiatives actually made a difference?
- Will you know which reforms were most effective?

This guide does not address all the ways in which you may be using ARRA and other funds to support educational reforms or all the approaches you could use to evaluate ARRA-funded and other educational initiatives in your state. Instead, the guide aims to help you think about opportunities for building evaluation into your ongoing efforts and define your evaluation priorities in two important ways:

1. **Framework.** The guide outlines a general evaluation framework that can help maximize what you learn about your educational reform initiatives. The framework consists of four basic questions that *can be applied to any type of program or initiative*. These questions can help you clarify your evaluation goals, ensure that you are set up to accomplish these goals, and know whether and how your initiative was effective.

2. **Examples.** The guide elaborates several sample evaluation questions on key ARRA topics to demonstrate how to apply the evaluation framework. Presented within distinct chapters, the examples focus on three ARRA priority areas: (1) increasing the effectiveness of teachers and leaders through professional development, (2) promoting the equitable distribution of effective teachers and leaders, and (3) turning around the lowest-performing schools. Many of the ideas explored in these examples can be extended to other reform efforts, including those that are not ARRA-funded.

By including both a general framework and concrete examples, the guide can help you consider the "big picture" of your evaluation efforts and also offer practical suggestions for how you can tackle your own evaluation questions.

Some of the reform initiatives in your state may be comprehensive—a combination of several interrelated programs or strategies—which can make it challenging to isolate the effects of specific components. Other new programs or reforms may be more narrowly focused. Each chapter of the guide therefore discusses two hypothetical examples—one focusing on a single initiative and the other focusing on a broader, interrelated set of initiatives.

Throughout the guide, our intent is to provide easy-to-use tools and resources—including tips and checklists—that you can use as part of an evaluation of most of the educational reform strategies that you are likely to initiate. These tips suggest ways in which you can reduce the need for special data collection efforts and make it easier to combine information from various data sources by building on existing data systems.

The guide also anticipates that you will work with an evaluation team to collect and analyze the necessary data and to prepare reports on the results. This evaluation team can involve a variety of individuals, including internal state or district staff and external experts who can be brought in to help with particular aspects of the evaluation. Designating an evaluation team will ensure that adequate resources are dedicated to the evaluation effort, that the needed technical and substantive skills are available to design and conduct the evaluation, and that the evaluation is at arm's length from program staff—an important factor to safeguard its credibility.

Why Do an Evaluation?

Why should you want to conduct an evaluation? Evaluations take time and resources, and you may think that there is already sufficient evidence that your programs are effective.

However, programs seldom operate exactly as planned and the context or circumstances in which they are implemented change. Anticipating that a program *should* help is not the same as seeing *how* it helps in your state. The information you acquire through evaluation can help you understand better the effect of *your* program on *your* schools, staff, and students. This information can help you refine your educational reform and improvement efforts. In addition, evaluations can provide information to others who are interested in your program, including state officials, target groups (teachers and other staff), administrators, and other individuals with a stake in the results (such as parents, students, and the general public). Different types of evaluations can help you learn different lessons about your programs and initiatives (see the text box below).

Evaluations can serve different purposes:

- ✓ ***Identifying service needs.*** "Needs assessment" evaluations can provide data on the knowledge and skills that everyone involved in your state's education system needs to acquire. This information can help you distinguish needs that can be met by existing services from those that will require new initiatives.
- ✓ ***Trying out a new program.*** The results of a "pilot" or "demonstration" effort—such as testing a new teacher compensation program in some districts or schools—can help you refine an innovative strategy before rolling it out statewide.
- ✓ ***Tracking program accomplishments.*** Keeping track of program activities—"implementation" studies—and documenting accomplishments—"outcomes" studies—can help you monitor progress toward goals and adjust your program to improve performance.
- ✓ ***Assessing whether the program was effective.*** Beyond determining whether particular program activities are being implemented as planned, an "impact" evaluation can help you determine whether the overall program is effective. For example, did teachers improve their classroom instruction more than they would have without the program? Did student achievement increase more than it would have otherwise? For which participants and in what contexts was the program most effective?
- ✓ ***Selecting among alternative approaches.*** A "comparative effectiveness" evaluation can help you identify and select the best practice among several options to achieve a particular goal, such as choosing among different approaches to turn around persistently low-performing schools.

A Framework for Thinking About Evaluation

The evaluation framework focuses on four questions that can help guide the reform efforts of evidence-based practitioners:

1. ***What are you trying to accomplish?*** It is important to clarify this question very early in the program development and evaluation planning process. Being clear about what you are trying to accomplish means thinking not only about the broad long-term goals of your initiative but also about how you expect specific activities and strategies to contribute to these goals. For example, the specific strategies that make up your

initiative may be linked and the success of one may depend on the success of another. Once you have a clear understanding of the initiative, you will be ready to specify short-term (or intermediate) indicators of progress for participating districts, schools, teachers, and students. For example, efforts to increase teacher effectiveness might include a short-term goal of increasing teachers' subject-matter knowledge or instructional skills to help meet the long-term goal of raising students' academic achievement.

2. ***How will you know whether your initiative was implemented well?*** Knowing to what extent the program was implemented as intended is crucial to understanding the results. Here are some important implementation questions:

   - Can you specify what successful implementation should look like?
   - Have you developed indicators of successful implementation, such as benchmarks for what needs to be accomplished at certain time points?
   - What data will you need to track implementation?
   - Do you have procedures in place to monitor implementation?
   - What barriers to successful implementation should you watch for?
   - What supports are in place to facilitate effective implementation?
   - How will you know whether these supports worked as planned?

   Answers to these and other questions are important to understanding program implementation. For example, if you are planning a new effort to help improve struggling schools, you could track which schools are targeted for assistance (and why), the existing conditions and what the schools have done in the past to improve the situation, what supports or assistance are provided and by whom, and the extent to which the schools (and the service providers) do what they are expected to do.

> **Implementation studies** involve monitoring the operations of a new program. Several types of questions are typically asked in these evaluations:
>
> ✓ ***Did the program (or a particular activity) occur as envisioned?*** If not, what barriers prevented it from being executed?
>
> ✓ ***To what extent were activities conducted according to the proposed timeline and scale?*** Was the program completed on time? Was the anticipated level of participation achieved? If not, what unanticipated obstacles were encountered?
>
> ✓ ***To what extent are actual program costs in line with budget expectations?*** Did the program cost more or less than originally planned? If so, why?

3. ***How will you track changes in outcomes?*** You might want to conduct a descriptive analysis to look for changes in the *outcomes* of interest for the participating districts, schools, teachers, or students. For example, after ARRA-funded initiatives have been implemented, you might want to know whether your state is moving toward a more equitable distribution of effective teachers and leaders across schools or districts.

Perhaps you will want to know whether teachers' overall effectiveness rankings have risen. Looking at changes in overall outcomes can be an important first step in building a theory about a program's results.

However, an outcomes study will not tell you whether the program caused the observed changes. Educational outcomes are influenced by many different factors acting simultaneously. For example, average achievement or proficiency rates may decrease in the schools participating in a new turnaround initiative, but this change may come about mainly because higher-performing students left the targeted schools. To estimate the ***effects*** of an educational reform or a new program, you need some way to describe what most likely *would have happened* in the targeted districts or schools if the changes had not been made. These evaluation activities are described next.

Outcome evaluations typically ask questions like these:

- ✓ ***To what extent are participants moving toward the anticipated goals of the program?*** Did key outcomes for participants show improvements? Is there evidence of changes in intermediary outcomes (such as student attendance or academic engagement) that may be important precursors to the desired final changes?
- ✓ ***Are the "gains" moving on the expected track?*** Are they smaller, larger, or about as expected?
- ✓ ***What factors, aside from the program, may be contributing to observed changes in outcomes?*** For example, have there been notable changes in the schools' leadership, staffing, or student composition? Have there been important changes in state or district educational policies?

4. ***Will you know whether your program was effective?*** Evaluations vary in the types of questions they can answer. Implementation and outcome evaluations provide ***descriptive*** answers to questions about *what* was done*, where* it was done*,* and *how* it was done, which can help you make sense of your implementation experiences. This information is useful and important, but it does not answer questions about ***effectiveness***—did your program make a difference? If you want to know whether your initiative was effective, you have to conduct an ***impact evaluation***. In impact or effectiveness evaluations, the intent is to understand not just what happened to the districts, schools, leaders, teachers, or students participating in the program, but *whether any observed changes would have happened if they had not participated*.

Impact evaluations require the ability to attribute changes in outcomes to the program or intervention being studied. The impact is not a correlation between program implementation and changes in teachers or students. It refers to *changes that would **not** have occurred absent the program or initiative being evaluated*. A few examples of impact questions are below:

- ✓ ***Was the program effective?*** Did the program **cause** a significant improvement in the desired outcomes?
- ✓ ***Was the program*** equally ***effective for all participants?*** Did some staff, schools, or student groups do better than others?
- ✓ ***What program*** components ***or activities were most effective?*** If different approaches or combinations of activities to achieving the same goal were tested, were some relatively more effective?
- ✓ ***What unintended*** impacts ***did the program have?*** Did unexpected or unintentional things happen?

Answering an impact question—were improvements in outcomes *caused* by your initiative?—requires additional advance planning. This is because *before* your initiative is implemented, you will need to define or identify a credible ***comparison group.*** A comparison group is a non-participating set of districts, schools, teachers, or students that provides information about *what most likely would have happened to participants absent the program or initiative being evaluated.*

Also important, the reliability of the answers provided by impact or effectiveness evaluations can vary, depending largely on the strategy used to set up these comparisons and the quality of the resulting comparison group (see Table 1). The chapters that follow present examples of different types of evaluation strategies you could use to separate the effects of your interventions, or of particular components of your interventions, from the effects of other factors.

Table 1: The Continuum of Rigor in Impact Evaluation Designs

Evaluation Design	Example	Type of Questions the Evaluation Can Answer	Rigor
Matched comparison group	Schools are selected to implement a new program through some nonrandom process (e.g., they volunteer). Before the program begins, these schools are matched on important background characteristics (e.g., student demographics and average test scores in the prior academic year) to other, nonparticipating schools. After the program has been implemented in the participating schools, the outcomes for the two groups of schools are compared to estimate the program's effect.	Did outcomes differ between the matched groups of participating and nonparticipating schools?	**Least Rigorous Design** (Lowest confidence that results can be attributed to program) ↓
Comparative interrupted time series	Schools are selected to implement a new program, again through a nonrandom process. Before the program begins, these schools are matched to comparison schools with similar ***histories*** of background characteristics and outcomes. After the program has been implemented in the participating schools, ***trends*** in outcomes over time are compared for the two groups of schools to estimate the program's effects.	Did outcomes in the program schools improve more than would be expected ***given trends*** in similar nonparticipating schools?	
Regression discontinuity	Schools are selected to implement a new program based on a predetermined "cut point" on a well-defined and easily measured criterion (e.g., proficiency rates below 25 percent). The outcomes for participating schools are then compared with the outcomes for schools that "just missed" being selected.	What is the impact of the program on outcomes? Or, are outcomes in program schools different than they would have been absent the program?	
Random assignment	A set of schools is selected to implement a pilot program based on a random process (e.g., a lottery is used to select 20 pilot schools from among interested volunteers statewide). At the end of the pilot implementation period, the outcomes for pilot schools are compared with the outcomes for the other interested, non-participating schools.	What is the impact of the program on outcomes? Or, are outcomes in the pilot schools different than they would have been absent the program?	**Most Rigorous Design** (Highest confidence that results can be attributed to program)

[This page was intentionally left blank for double-sided printing.]

Evaluating Strategies to Promote Teacher and Leader Effectiveness Through Professional Development

by Michael Puma, Michael Garet, Kirk Walters and Jessica Heppen

> *[Our] ongoing education reform approach is built on the knowledge that the surest avenue to improving student achievement is improving and supporting the practice of instructional leaders and teachers.*
>
> *Alabama Dept. of Education (2010, p.80)*

States and districts can improve staff effectiveness by using a range of strategies: improved pre-service preparation, targeted recruitment efforts, new-staff induction programs, professional development opportunities, enhanced links between compensation/incentives and performance evaluations, and leadership development.[1] To illustrate how you can incorporate evaluation into your new programs and initiatives, this chapter focuses on one key strategy that is an important part of many staff effectiveness efforts: **professional development** (PD) targeted to teachers and instructional leaders.

As discussed in the Introduction, it is best to think about evaluation when you are developing your plans for a new program. This way, you can be sure to have the information you need to answer important questions later. It also can be easier and more efficient to build in data collection from the beginning. Two hypothetical examples illustrate how each step of the previously described evaluation framework might apply to a PD initiative that a state might actually implement.

1. **Data "dashboard" system.** This example is a relatively low-intensity PD program designed to help teachers and principals more easily use data to improve reading and math instruction. The example focuses on increased use of data by all relevant subject-area teachers and their schools for instructional decision making. In this example, the training is provided by specialized trainers directly to all targeted staff. The pedagogical changes that are expected to occur, such as using data to improve instruction teachers' instruction, are generally specified by the PD program.

2. **PD academy for differentiated instruction.** This example is a more-intensive PD program that targets a subset of teachers and leaders who work in the lowest-achieving schools in the state. At the end of the PD program, the teachers and leaders who have been trained are expected to take back what they have learned to help other staff in their respective schools implement differentiated instruction—identifying the learning needs of individual students and tailoring instruction to meet those needs. Unlike the first example, this model uses a "train the trainer" approach. The pedagogical changes that are expected to occur, such as greater use of flexible grouping and differentiated materials, are highly specified by the PD program.

These two examples address different types of evaluation issues that state officials might encounter. This discussion assumes that you will work closely with your districts and schools

[1] National Comprehensive Center for Teacher Quality, 2009.

to implement the reform initiatives. It also expects that you will work with an evaluation team to collect and analyze data and to prepare reports on the results. This cooperation will ensure that adequate resources are dedicated to this purpose, the necessary skills are available, and that the evaluation is at arm's length from program staff. The evaluation team can consist of state staff, external experts, or a combination of both.

Example 1: Helping Practitioners Use Data to Improve Instruction

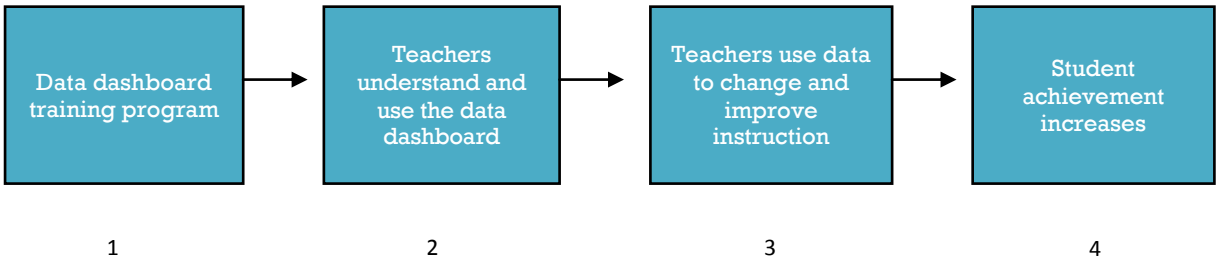
A state has been using student achievement gains to evaluate teacher performance since the 1990s. As a tool to help improve instruction, the state recently granted access to these data to all teachers in the state. However, the state knows that providing access to the data won't automatically lead to improved teaching and learning. So, the state plans to work with districts to train and support teachers and administrators on how to make the most of this information, which is presented as an online ***data dashboard*** *system. Education dashboards provide school- or classroom-level information in a concise way that can support instructional decision making. The information typically includes student test score data but can also include other important indicators such as grade promotion, graduation rates, and attendance.*

The training, to be conducted by the districts with the help of a nonprofit partner, will be provided to teachers during the 2010–11 school year. As training modules are completed, they will also be made available online for teachers who missed the training and for teachers to access as ongoing needs arise.

1. What is the state trying to accomplish through this PD initiative?

The ultimate goal of this initiative is for teachers and principals to use data in ways that increase student learning (Figure 1). Although focusing on this ultimate outcome is clearly important, the state should also think about how receiving the planned PD is expected to lead to higher student achievement. That is, what are the participants expected to learn, how will they use the knowledge and skills back in their schools, and how is that use expected to affect student outcomes? *Why is this important?* If teachers' use of the data system does not immediately lead to improved student outcomes, an exclusive focus on student achievement won't let you know where to look to understand why the initiative was or was not successful.

Figure 1: How Is PD Expected to Improve Student Outcomes?



As shown in Figure 1, this initiative assumes that the planned PD can help teachers understand the information made available to them and how to use it improve their instructional practices. For example, data on the achievement gains of individual students could help them better target their instruction, especially to those students who the data show are struggling with specific skills or particular concepts. In turn, teachers are expected to incorporate these data-driven instructional techniques into their classrooms, and as a result, student achievement is expected to improve. Given this anticipated chain of events, you may find it helpful to track these types of intermediate outcomes (results that need to happen to reach the ultimate goal) and to use them as early indicators of progress.

Laying out these intermediate steps may also help you set a more realistic timeline for when to expect changes in student achievement. That is, it will take time for teachers to gain proficiency in using data and to learn how to use the data to alter their instructional practices. These intermediate changes in knowledge and practice must be sufficient to cause measurable improvements in students' learning.

Clarifying intermediate outcomes might help you set a more realistic timeline of when to see final results.

2. How will the state know whether the initiative is implemented well?

It is also important to monitor implementation so that you will know the nature of the PD that teachers ultimately receive. Monitoring implementation can be costly and time-consuming, but it does not need to be. Here are some simple things that you can do to track the implementation of a PD program.

PD IMPLEMENTATION COMPONENTS TO TRACK

- ✓ **Who** delivered and participated in the PD?
- ✓ **How** was it delivered (type and duration of sessions)?
- ✓ **When** did the PD occur?
- ✓ **What** was the focus of the PD session(s)? What materials were used?
- ✓ **Satisfaction:** Were the participants engaged? Did they find the PD useful?

First, the trainers could be required to distribute and provide the evaluators with **sign-in sheets** at all PD events, including the centralized trainings for trainers and the sessions between the trainers and teachers. If you also use online training, be sure to electronically capture its use as well. Making sure that you know who attended, how often, and for how long is easy. It is also important for understanding whether the program was implemented as planned and at a scale that addresses a sufficient number of teachers across the state. To evaluate implementation in more detail, the evaluators could also **collect the training agendas** used across sites and trainers and review them for consistency and alignment with the intended content.

Something that will require more effort, but may be worth considering, is to have someone from the evaluation team **attend a sample of training sessions** to keep track of whether the topics in the agenda were covered and how closely the planned agenda matched what was actually delivered. The evaluators could also **check in with the trainers** to get a sense of how things are going. This check could be as simple as a quick conversation at the conclusion of the training. Or, you could create an email template that trainers return after each PD event that describes any adjustments they made (or will make in the future) to the training agenda.

If coaches are part of the initiative, it may also be helpful to gather data from them about how they are allocating their time. An easy approach is to have the coaches record their activities in a simple **coaching log** that will enable you to know what they focus on, and for how long, as they work with participating teachers.

Finally, you will probably want to collect **participant feedback** at the end of each training session. This can be easily done by using a short post-training questionnaire to get trainee feedback on how well the

SIMPLE WAYS TO TRACK PD IMPLEMENTATION

- ✓ Distribute and catalog participant sign-in sheets.
- ✓ Obtain trainee feedback.
- ✓ Develop training agendas to keep track of how closely the training matched the plan.
- ✓ Have coaches keep a log of activities with teachers.
- ✓ Check in with the PD trainers and coaches along the way.

training went, whether they found it useful, and if they expect to use the skills learn back in their schools.

These are just a few things that will help you track how the dashboard PD is being implemented. If your evaluation team wants to look at some sample implementation instruments before developing its own, several from other studies are available to the public. The resource link below includes examples of PD fidelity forms and coaching logs that offer some ideas about where to start.

RESOURCE LINK: MEASURING PD IMPLEMENTATION

- ✓ Sample PD Implementation Forms and Coaching Logs:
  - PD Impact Reading Study: http://ies.ed.gov/ncee/pubs/20084030/
  - Middle School Mathematics PD Impact Study: http://ies.ed.gov/ncee/pubs/20104009/

3. How might the state track changes in outcomes?

Once the training has been launched and implementation data are being collected, you can start tracking the outcomes of interest. For this PD model, you may want to track outcomes related to teachers' understanding and use of the dashboard system, teachers' instructional practices, and student achievement.

- ***Teachers' understanding and use of the data dashboard system.*** A fairly straightforward approach to assessing participants' **understanding** is to have your evaluation team collaborate with the training provider to develop an assessment of teachers' knowledge of how to operate the system. Teachers could take the assessment at the conclusion of the training, with the forms being collected by the trainers and provided to the evaluator for subsequent analysis.

  Because the data system in this example is automated, the state could easily capture teachers' **use** of the system electronically, such as the number of minutes per week that they access specific parts of the dashboard. A somewhat more resource-intensive approach would entail periodic surveys of participating teachers to collect information on whether, and how, they used what they learned once they were back in their schools. This survey can be done through an online system, if the evaluators have this capability, or through paper-and-pencil questionnaires that are distributed to staff through their schools. Another, more resource intensive approach is to conduct focus group interviews with selected participants after they are back in their home schools.

- ***Teachers' data-driven instructional practices.*** As noted above, the training is intended to help teachers understand how they can use the dashboard data to improve their instructional practices. For example, a math teacher could use the data to see which skill areas are creating difficulties for her students and then alter how these skills are taught and/or spend more instructional time on those skills. Of course, measuring such changes is challenging. One approach could involve identifying particular teacher behaviors that you expect to see happen more frequently as a result of the training, such as within-class ability grouping and greater use of small-group instruction and individual tutoring. You could then incorporate

these items into the observation protocols that principals may use currently (or be planning to use) as part of their teacher evaluation processes. Alternatively, your evaluation team could develop a tailored protocol for their own classroom walkthroughs for a selected sample of participating teachers.

- ***Student achievement.*** In this example, the state is already tracking student outcomes. It can use these data to get an initial look at progress being made across the state by comparing test score trends before and after the PD training at individual schools. The state could also compare such student achievement trends within groups of similar schools, contrasting trends for schools that have and have not received the training. It is important to keep in mind, however, that this method does not tell you whether the PD "caused" any observed changes in test scores, an issue discussed in more detail below.

4. How might the state determine whether the training program is effective?

Just knowing whether outcomes have improved doesn't tell you that it was this program that made a difference. To truly assess program effectiveness, you need a fair and realistic group of nonparticipating schools (or teachers) with which you can compare participating schools (or teachers). The goal here is to determine whether the students of teachers who participated in the dashboard training do better than they would have if their teachers had not received the training. Answering this question is challenging, but less so if you have a thoughtful strategy for rolling out the program.

Often, states seek to provide a new program to all eligible districts, schools, or teachers at the same time. Although this makes it possible to fully implement the program quickly, it is beneficial to consider staggering implementation so that some schools receive the program right away and others receive the program a year (or more) later. Taking this approach can allow you to evaluate the program's effectiveness because the schools that do not participate initially can serve as a comparison group for the schools that do get the program from the start. Some objections may be raised because of the perceived "withholding" of services from some schools or teachers. However, it is often the reality that it is infeasible to implement such a program all at once statewide, and there is no actual denial of benefits, just a planned phased implementation in which everyone will be served. In addition, this approach allows you to learn as you go and to make any necessary adjustments to the program.

To truly assess program effectiveness, you need a fair and realistic comparison group of nonparticipants.

Take this example of the data dashboard PD program. If the state delivered the program to some portion of the schools in the 2010–11 school year (Year 1) and to the rest of the schools in 2011–12 (Year 2), these two groups of schools could be compared on key outcomes to determine the program's effectiveness. Table 2 illustrates how this would work.

Table 2: Staggered PD Implementation Design

	Year 1 2010–11	Year 2 2011–12	Year 3 2012–13	Year 4 2013–14
Initial trainees	Use dashboard system with training	Use dashboard system with access to online modules	→	
Delayed trainees	Use dashboard system without training	Use dashboard system with Training	Use dashboard system with access to online modules	→
Analytic Activities	Evaluate and refine Year 1 training	Refine Year 2 training ---------- Track outcomes of both cohorts	→	

In this example, the initial trainee group would use the dashboard system and also receive the training, while the delayed trainee group would have access to the dashboard system but would not receive the PD training right away. At the end of Year 1, the two groups could be compared on key outcomes to determine the effectiveness of the PD program. In Year 2, the delayed trainee group would receive the same training that the initial trainee group received in Year 1. Both groups would have access to the online training modules as they continue to use the system in Years 3 and 4. The evaluation team could continue to track outcomes of both groups of schools in Years 2 to 4.

This example assumes, for illustration purposes only, that the observed effects of the dashboard training could be seen after a single year. As noted earlier, however, this expectation may be unrealistic. Hence, you may want to consider the important trade-off between staggering implementation to allow enough time for the PD initiative to show effects against the desire to have most schools implementing the new program as soon as possible. A longer delay in providing the dashboard training to the comparison schools may improve your ability to observe changes in trainees' data use and instruction as they gain more experience with the concepts taught in the training.

The key for this staggered implementation design to support your evaluation goals is that it helps you establish **equivalent** groups of schools to compare. How might you do this? You have several options, but the following two may be most feasible.

- ***Use a lottery to create equivalent groups of schools.*** You would start by identifying the pool of schools that is intended to benefit from the PD program, such as all elementary schools, all secondary schools, all schools that didn't meet a specific performance target, or all schools. Then, you would decide on the rollout schedule; say, half the eligible schools would receive the training in Year 1 and the other half in Year 2. The third step would be to use a random process to assign schools to each group. This process could be as simple as listing all the schools in an Excel spreadsheet and using the included random number generator to assign a number between 0 and 1 to each school. Schools with values under .5 could be designated the Year 1 schools, and schools with values over .5 could be the Year 2 schools. The two groups of schools would then be statistically equivalent, which means that there should be no systematic differences between them before either group received the training. Comparing their outcomes at the end of Year 1, or in later years, would

provide a credible measure of the impact of the PD program. As noted in the Introduction to this guide, this is one of the strongest designs available for determining program effectiveness.

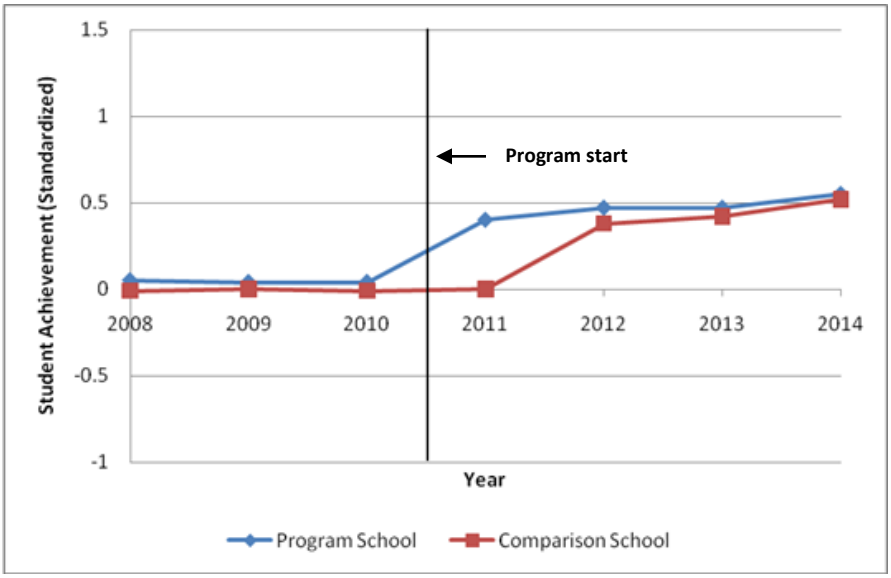
- ***Establish matched comparison schools.*** If using a lottery isn't possible, the schools selected to receive the training in Year 1 could be chosen in some nonrandom way. For example, you could match the schools selected for the initial training with schools that are most like them on a set of available characteristics. These characteristics can be at the school level, such as grade levels served, enrollment size, and student-staff ratios. They can also be at the student level, such as eligibility for subsidized school meals, average proficiency levels on state assessments, attendance, grade promotion, and graduation. With this approach, your goal is to establish a group of comparison schools that is as equivalent as possible to the schools being trained.

  Here, having multiple years of prior information, such as trends in achievement, enrollment, mobility, and demographic characteristics, is essential. To establish a group of schools that can serve as a fair comparison with the group selected to receive the training program in Year 1, you would need to ensure that the expected performance trends into 2010–11 would be the same for both groups in the absence of the program; that is, without the training program, you would expect the two groups of schools to perform similarly during the 2010–11 school year. Once the matched comparison schools are established, the two groups can be compared on outcomes at the end of Year 1 (and later) to assess effectiveness of the program.

  
  Multiple years of background information help make the matched comparison group as equivalent as possible to the treatment group.

  Figure 2 shows a hypothetical example in which two groups of schools—the program group that received the dashboard training and the comparison group that did not—are compared over time. As shown, the two groups of schools performed similarly, on average, during the three school years before the training program was implemented (2008, 2009, 2010). In 2010, the dashboard PD training was delivered to staff in the program schools, and in this example, there was a **positive effect**. Program schools performed better in 2011 than comparison schools. Then in 2011, the comparison schools received the program. By 2012 and 2013, the schools were performing similarly to the original program schools, which is also better than how all the schools performed before the program.

Figure 2: Hypothetical Matched Comparison Results



Example 2: Summer PD Academy on Differentiated Instruction

To improve learning at the lowest-achieving schools, a state has planned an intensive summer academy to provide PD on differentiated instruction to lead teachers, instructional coaches, and department chairs from the state's persistently lowest achieving schools (defined as the lowest 5 percent of schools ranked on student test scores). Trainees are expected to return to their respective schools and train the remaining teachers and administrators—often referred to as a train-the-trainer model.

Differentiated instruction recognizes students' varying background knowledge, readiness to learn, language, interests, and motivation. It seeks to tailor classroom instruction in a way that maximizes each student's growth and success in school. Teachers learn how to tune in to the individual needs of their students and adapt their instructional practices accordingly. Examples of differentiated instruction strategies are flexible grouping (whole group, small group, pairs, and individual tutoring), continuous assessment of student progress, student choice of learning opportunities, and differentiated instructional materials.

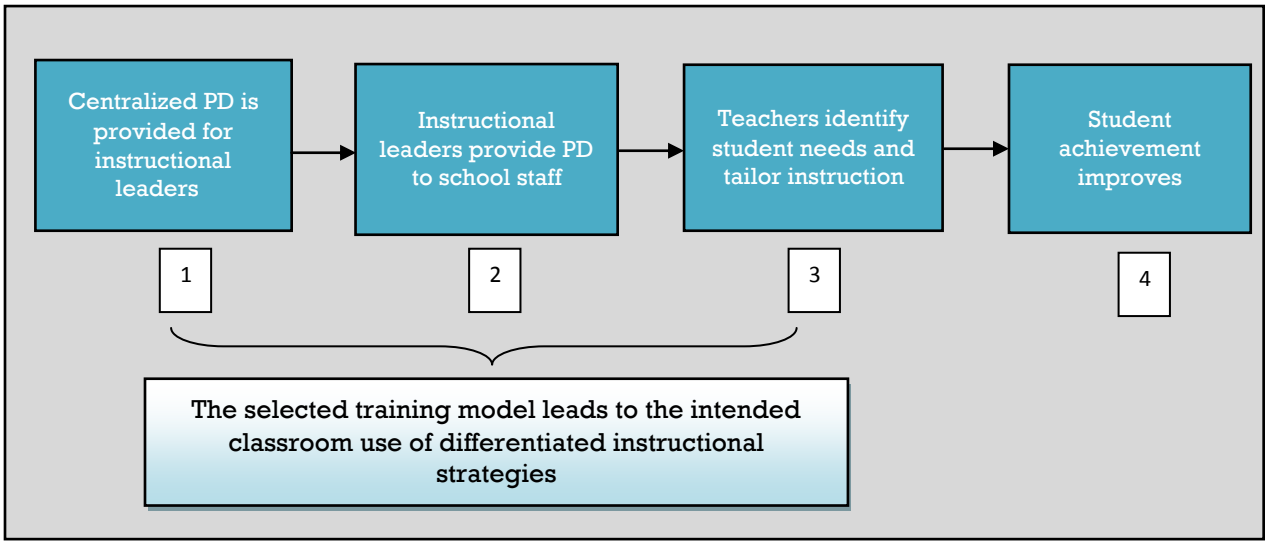
Unlike in the first example, specific instructional strategies are part of the training, which will also vary by subject and grade. That is, appropriate instruction in math can be different from pedagogical approaches that work best in reading, and what is best suited for children in the early elementary grades may be quite different from the best way to teach older students. These factors need to be considered in planning for evaluation.

Although both examples in this chapter focus on PD initiatives, this PD example is different from the previous example in several respects. First, PD will be provided to teachers indirectly, through a train-the-trainer model, which is different from the direct approach used in the data dashboard example. Second, the PD is about specific instructional practices, so the expected changes in classroom practices are explicitly prescribed as part of the training, making the anticipated intermediate outcomes easier to define. Finally, the context in which PD is being delivered focuses on teachers and students in schools that are persistently low achieving and under intense pressure to improve. These differences have implications for the focus and approach of an appropriate evaluation.

1. What is the state trying to accomplish through this PD initiative?

The state is using a train-the-trainer approach to scale up a set of instructional practices that it expects will help teachers in low-performing schools improve student learning. Because the targeted schools are in danger of being closed down or taken over by the state if they do not show academic improvement, the PD has an added importance to participants. This state's change model might look like the one in Figure 3.

Figure 3: How Is PD Expected to Improve Student Outcomes?



Similar to the previous example, it would be a good idea to identify intermediate outcomes to evaluate as the program is being implemented. In this case, these outcomes would include teachers' ability to identify each student's individual learning style and needs, the use of differentiated instructional strategies, and the appropriate matching of student needs to instructional strategies. As in the first example, thinking about these intermediate goals may influence your implementation timeline and expectations related to the initiative's ultimate goal: improving student achievement in persistently lowest-performing schools. Unlike in the previous example, however, you also have to consider the expectation that the certified trainers are able to appropriately train their school peers, the time it will take for this more complex set of strategies to be internalized and applied by the teachers, and the time needed for this multidimensional reform to result in improved student outcomes.

2. How will the state know whether the initiative is implemented well?

In this example, you may be interested in monitoring not only the implementation of the summer academy but also the school-based PD subsequently provided by the certified trainers to their peers. For the training that takes place at the summer academy, the evaluators can use the same kinds of data collection strategies discussed in the first example: sign-in sheets, training agendas, and session observations.

Monitoring the school-based implementation is likely to be trickier because the certified trainers are likely to vary the form and intensity of PD they provide to other teachers at their schools. The PD is also likely to vary by subject and grade. For example, they may work with different-sized groups of teachers and place more or less emphasis on various differentiated instructional strategies. They might work with individual teachers in a coaching model, opt for a single training session early in the school year, or spread training out over the entire year. Consequently, you may want to identify the core training activities that should occur in all schools and have your evaluation team collect data on whether these happen in the participating schools. Such data could be collected through a simple teacher survey or through observations of teacher training activities in a selected sample of targeted schools. You shouldn't worry about trying to capture all the possible variations. Instead, it

probably makes more sense to determine whether, in general, the program is being implemented as expected.

3. How might the state track changes in outcomes?

After the PD has been implemented, you can begin to track intermediate and final outcomes. Here are some examples:

- ***Instructional leaders' certification rates.*** In this example, a relatively easy step would be to establish a certification process as part of the academy. The next task would be to collect data on the percentage of trainers who pass the certification exam and the number of attempts they needed to become certified.

- ***Teachers' knowledge and practices.*** In this example, the training program is expected to emphasize a variety of teacher skills, such as the ability to identify student needs and use particular instructional strategies. How the development of these skills plays out will likely vary by grade and subject. Consequently, measuring the breadth of possible changes can be challenging, even for experienced evaluators, and can require a large commitment of resources. As an alternative, you might consider modifying an existing (or planned) teacher evaluation protocol to add some key indicators that could help principals and your evaluation team determine how well teachers have incorporated the basic aspects of differentiated instruction into their instructional routines.

- ***Student achievement.*** In this example, the state has used student achievement data to identify the target schools as persistently low performing. Assuming that the state does not change the assessment system and continues to assess schools' academic performance as it did in the past, the state will be able to easily track whether student achievement changes in the years after introducing the summer academy PD program. The state could also break out achievement trends by grade level and subject area to get a more detailed picture of changes in academic achievement.

To take a snapshot of the implementation of school-based PD, the state might want to identify core training activities to be delivered in all schools.

4. How might the state determine whether the training program was effective?

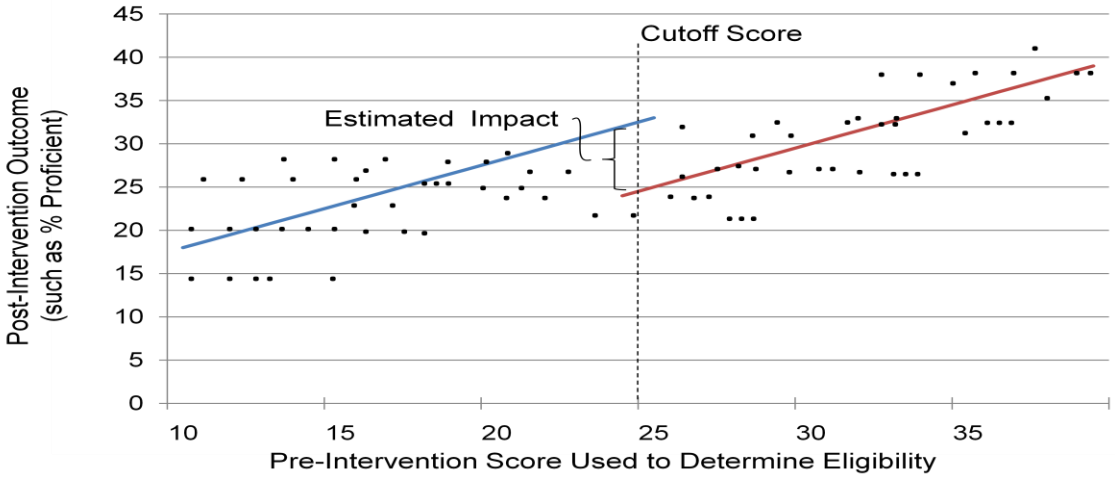
Although tracking changes in overall outcomes is clearly an important step, ultimately you will want to know whether any changes in outcomes over time are due to the summer academy. As with the previous example, if you decide to stagger the implementation of the program, you could establish equivalent groups of schools by using either a lottery or a matched comparison strategy like those discussed earlier.

Another approach for assessing whether the program is effective takes advantage of the fact that, in this example, the state decided to target the training academy to the lowest-performing 5 percent of schools in the state. By targeting schools (or teachers or students) on the basis of objective, quantifiable indicators that occurred before the program begins, comparison groups of schools can be established. This means that schools with the highest need can be determined eligible for the program, and you can still assess the effectiveness

of the program by comparing them with noneligible schools.[2] The basic idea of this design is that the eligibility cutoff score is used to assign schools (or teachers or students) to either the participating (program) group or the nonparticipating (comparison) group. Schools on one side of the cutoff get the program and schools on the other side of the cutoff do not. In this example, the cutoff score is the measure of school-average student achievement that marks off the lowest-achieving 5 percent of schools from the rest of the schools in the state. (In some cases, it may be best to define such cutoffs separately for elementary, middle, and high schools, which is fine as long as the rule for assigning schools to the program is not broken.)

Figure 4 displays a graph that shows hypothetical results from an evaluation design set up this way. The eligibility measure was defined as the average percentage of students per school who scored at or above proficient on the state test in spring 2010. The cutoff score was established at 20 percent, suggesting that in this state, the lowest performing 5 percent of schools are those where fewer than 20 percent of students score at or above proficient. The schools that participated in the training academy, or program schools, are represented by the dots on the left side of the cutoff in the graph. The nonparticipating comparison schools are the rest of the schools in the state, represented by the dots on the right side of the graph. The blue and red lines in the graph are fitted regression lines that describe the relationship between the pre- and postprogram proficiency rates that might be observed in each group of schools. These hypothetical results suggest a positive effect of the program that is equivalent to approximately 8 percentile points on postintervention proficiency rates. The effect can be observed right around the cutoff line.

Figure 4: Hypothetical Regression Discontinuity Results



Note: This figure uses hypothetical data and regression lines to demonstrate a regression discontinuity design with an outcome such as proficiency rates on the Y-axis and the score used to determine eligibility for the intervention on the X-axis.

Evaluating programs in this way can be attractive because it does not require assigning schools in need of support to a nonparticipating or delayed-participation comparison group. When using a cutoff score to establish participating and nonparticipating groups, however, you need to consider other factors:

[2] This is formally called a "regression discontinuity" or RD design.

- **Defining the preprogram eligibility measure.** The preprogram measure used to establish the cutoff must be a quantitative measure that can be measured on a continuous scale. Measures of achievement or school poverty can be useful preprogram measures to establish the cutoff.

- **Defining the cutoff.** Once you have defined the preprogram measure, the choice of a cutoff point depends on your goals for implementation, including the resources available to fund the program. A state may know that the program can be delivered to a set percentage of schools and would then need to decide how to separate the target schools from the rest of the schools.

- **Strictly using the cutoff.** All schools below the established cutoff are assigned to the program group, and all schools above the cutoff are assigned to the nonparticipating group. Any exceptions can distort the estimates of program effectiveness.

- **Having a sufficient number of schools.** This type of evaluation design requires more schools than do evaluations that use a lottery to establish groups of equivalent schools. In statewide evaluations, this should not be a problem, but in all cases, determining whether the number of schools is sufficient to detect program effects is essential. Your evaluation team should carefully consider this issue.

Evaluating Initiatives to Promote the Equitable Distribution of Effective Teachers

by Irma Perez-Johnson, Samia Amin, and John Burghardt

States and districts are adopting a range of strategies to help ensure that all students have equitable access to teachers and school leaders with strong track records in helping students learn. These strategies can include policies and practices to recruit, hire, place, and reward effective teachers and leaders, to remove ineffective ones, to improve teachers' working conditions, and to strengthen their professional skills. Many of these strategies target teachers' decisions about where to teach and whether to remain there over time.

Some states and districts structure their equitable distribution programs to attract highly effective teachers to schools considered high need because of their location, proportions of low-income or otherwise disadvantaged students, and/or other characteristics. To illustrate how you can use evaluation to assess and refine equitable distribution initiatives, this chapter focuses on teacher compensation and related policies designed to help attract and keep effective teachers at high-need schools.

This chapter applies the evaluation framework described in the Introduction to two examples of equitable distribution initiatives:

1. **A tax-exempt signing-bonus program**. This example involves a state-funded program that provides additional resources to high-need schools so that they can offer tax-exempt signing bonuses to experienced, effective teachers who agree to transfer to these schools.

2. **A teacher recruitment, retention, and improvement (TRRI) initiative.** This example describes a multifaceted program in which high-need schools adopt a range of compensation strategies designed to help attract, retain, and motivate effective teachers, coupled with targeted professional development (PD) and mentoring to enhance these teachers' skills when working with high-need students.

These examples illustrate how evaluation methods can be applied both to discrete teacher compensation initiatives and to multifaceted programs that mix financial incentives with other supports for teachers working in high-need schools. ***Many aspects of these evaluation examples are relevant to evaluating other promising reform efforts that use other strategies to promote the equitable distribution of effective teachers.***

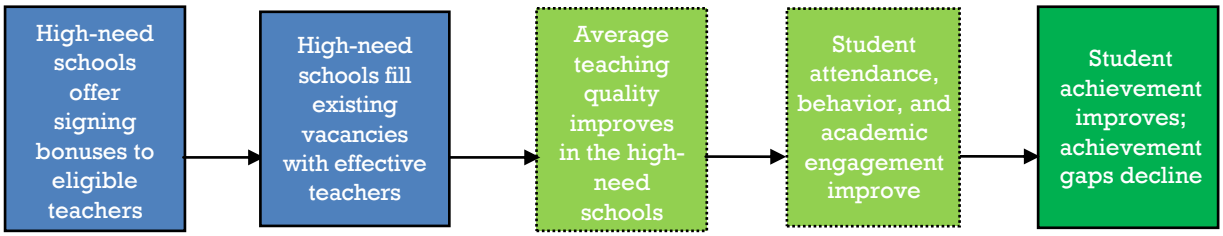
Example 1: A Tax-Exempt Signing-Bonus Program

The state is implementing a program that enables high-need schools to offer signing bonuses as they recruit high-performing teachers. Qualifying schools in the state are authorized to offer signing bonuses of up to $50,000 vested over three years, up to a statewide total of $10 million. The bonus is exempt from state taxes. It is also contingent on the teacher demonstrating continued effectiveness when working with students at the high-need school.

1. What is the state trying to accomplish through this initiative?

Teacher compensation policies to promote the equitable distribution of effective teachers ultimately aim to improve student learning at targeted schools and to reduce gaps in student performance between targeted schools and other schools. However, it is also important to identify and track other outcomes that may signal change and become evident sooner than improvements in student achievement. Figure 5 shows some outcomes that might be intermediate benchmarks of progress for the signing-bonus program in this example. (Note that this chapter builds on this basic chain of events, or change model, for the signing-bonus program later on.)

Figure 5: A Basic Change Model for the Signing-Bonus Program



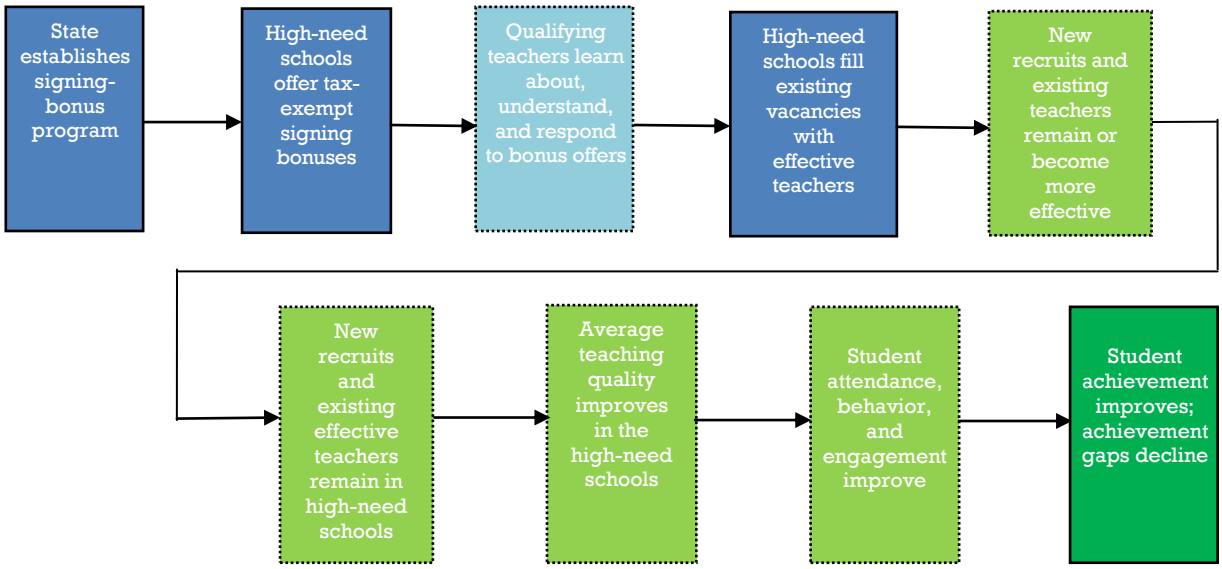
Mapping the sequence of steps expected to lead to improved student achievement can help clarify your assumptions about how program elements will fit together and interact to produce the anticipated outcomes. You might decide that some of these assumptions have adequate supporting evidence, whereas others may need closer examination as the program is implemented. For instance, in this example, a key assumption that may merit attention is that the ***changes in average teaching quality*** in the high-need schools will be large enough to influence student outcomes. This outcome may depend on how many vacancies at the participating high-need schools are filled with qualifying teachers, whether the newly recruited teachers remain effective and continue teaching at these schools, and how the new teachers interact with or influence other teachers in the targeted schools.

Thinking through the program's change model can also focus attention on how long it may take for the anticipated changes to occur. This is important for establishing realistic time lines for program design, implementation, and data collection. For instance, ***changes in student attendance, behavior, and overall academic engagement*** may be precursors to improvements in academic achievement, such as test scores. Data on the early steps in the sequence may also indicate a need to modify the program. For example, if the signing bonuses fail to attract teachers with the hoped-for qualifications, you would be able detect this

early on and change the program publicity, the terms of the bonus offer, or other aspects of the program to make it more attractive to the targeted teachers.

Developing a more detailed change model for the program could help you translate program assumptions into measurable indicators of progress. Figure 6 expands the basic change model to incorporate possible additional assumptions.

Figure 6: A More Detailed Change Model for the Signing-Bonus Program



As Figure 6 shows, the signing-bonus initiative assumes that effective teachers at other schools will learn about and understand the financial incentive being offered by the high-need schools and that they will respond to the offer. It further assumes that effective teachers who transfer to the high-need schools will continue to be effective, even though the background characteristics of the students they teach and the environment in which they operate will have changed. It also assumes that not only new recruits but also capable teachers already in place will choose to remain at the targeted high-need schools. Tracking the mobility of new and existing teachers and interim outcomes such as changes in teacher instructional practices or the academic climate in the targeted schools can help you examine program assumptions more closely.

2. How will the state know if the initiative is implemented well?

Tracking the intermediate and final outcomes of the signing-bonus program is clearly important. However, these outcomes are likely to occur only if the program is implemented well and has no unintended negative consequences in other areas, such as the decisions of other effective teachers already teaching in the targeted schools. When a new program like this signing bonus is introduced, you may want to track implementation closely and see how it actually plays out. The resulting information can help you get a sense of how likely the anticipated benefits are to materialize. Knowing which elements of the program unfold as planned and prompt the hoped-for responses, and which do not, can also provide important lessons for future programs.

For a signing-bonus program, for example, you will need to decide which schools to target and which teachers are eligible for the program. You might want to track how well these targeting policies work as the program is implemented. These decisions might be formulated as follows:

- **Defining high-need schools.** You will need to decide which schools are eligible to receive bonus program funds to offer to prospective hires. States typically designate schools as high need on the basis of several factors, such as ***low academic performance*** (based on average standardized test scores compared to those of similar schools); ***high poverty*** (based on the percentage of students eligible for free or reduced-price lunch); ***remote or rural status*** (based on geographic location); and ***minority status*** (based on the percentage of students from minority families or students who are English language learners). Instead of relying on a single factor to designate schools as high need, you could combine schools' "scores" or rankings on several indicators of need and create a composite measure that takes into account multiple factors. You could also choose to give some criteria, such as academic performance or poverty, more weight than others in this measure.

- **Identifying effective teachers.** The state must also define the criteria for effective or high-performing teachers who can receive a signing bonus. Teacher effectiveness can be defined in a variety of ways. ***Student outcomes*** are likely to be an important element of your measure of teacher effectiveness, and there is increasing emphasis on relying on value-added measures for this purpose (see text box below). You may also include assessments of teachers' ***instructional practices*** in your measures of effectiveness. These assessments may be based on classroom observations, peer review, principal ratings, or a combination of these, guided by rubrics such as Danielson's Framework for Teaching (2007) or the CLASS (Classroom Assessment Scoring System; Pianta, LaParo, Hamre, 2008). Credentials (basic or advanced certification, advanced degrees), experience (years in teaching), and background knowledge (SAT/ACT scores, grade point average) have traditionally served as indicators of teacher effectiveness, although they have been shown to be poor predictors of student learning. You may nevertheless decide to continue to use such measures as you develop or refine effectiveness measures that are based on student growth, teacher practice, or both. Again, you may opt to combine multiple indicators of teacher effectiveness into a composite measure, rather than rely on a single indicator.

Evaluating Teacher Effectiveness Using Value-Added Models

Assessing teacher effectiveness is clearly challenging. Traditionally, schools and districts have relied on credentials such as educational background, certification, experience, or advanced degrees for this purpose. **However, credentials have been shown to be poor predictors of how much students learn with a given teacher.** For this reason, some states and districts have been experimenting with **value-added models** that directly estimate how much individual teachers contribute to the learning of students in their classrooms. Important considerations regarding the use of such models include the following:

- ✓ To be reliable and fair, value-added models must account for what students already know, differences in the background characteristics of students assigned to different teachers, and other important influences on student learning.
- ✓ A value-added approach cannot be applied to all teachers because achievement tests are not administered in all grades or subjects. It is also difficult to isolate individual contributions when teachers team-teach or provide supplementary instruction. Alternative evaluation methods must be developed in such situations.
- ✓ The stability and accuracy of the **estimates** of teacher effectiveness generated by these models must also be considered, especially when they are used for "higher stakes" policy decisions such as tenure, dismissal, or financial rewards.

After identifying teachers who are eligible for the signing bonus, the next steps are to publicize the program to them, ensure that they understand the bonus offer, and encourage them to respond to it. Hence, the success or failure of the program may be influenced by other factors, such as the size and complexity of the bonus offer, the conditions for receiving the bonus, and the effectiveness of publicity efforts. If program implementation falters, you will want an accurate diagnosis of the barriers encountered before you invest further resources or make program adjustments. For example, increasing the bonus offer amount may not help if teachers are unaware of the program or consider the conditions for receiving the full bonus too difficult to meet.

The text box below lists some implementation issues that states adopting signing-bonus programs like the one described here may want to examine. The list is not exhaustive, but it can be a useful starting point. It also highlights some of the sources of information that you may be able to tap to monitor implementation. As these examples illustrate, collecting data on program implementation may require fewer resources than you anticipate, especially if you use existing administrative records infrastructures and strategically enhance them to collect desired implementation data.

Implementation Checklist

- ✓ **Targeting:** What criteria are used to determine which schools qualify for state funds to offer signing bonuses? How well do they align with high-need criteria? What criteria determine which teachers can be offered the bonus? How are these criteria established? Who determines whether a particular teacher meets the criteria? How is this done? **(Answer these using → administrative records)**
- ✓ **Bonus design:** What are the amount, duration, and conditions for payment of the bonus? To what extent and how do these features differ across districts, schools, and individual teachers? **(Answer these using →administrative records)**
- ✓ **Publicity:** How do teachers learn about the bonuses? Which publicity efforts are more successful than others in eliciting teacher participation? Do any sources of misinformation deter bonus uptake? **(Answer these using → teacher surveys or focus groups)**
- ✓ **Participation:** Which schools participate in the program and to what degree? Which teachers apply for and accept bonuses to transfer to high-need schools? **(Answer these using → administrative records, teacher surveys)**
- ✓ **Spillover effects:** What effects does the signing-bonus program have on the morale and retention of existing teachers in the high-need schools? What effects does it have on staff and students in the "source" schools? **(Answer these using → teacher surveys, administrative data on student performance and teacher mobility, measures of teacher performance)**

3. How might the state track changes in key outcomes?

Tracking the outcomes of the signing-bonus program will likely involve several steps, reflecting the steps in the program's change model (see Figures 5 and 6). Key outcomes that you might track include the following:

- **Changes in the ability of high-need schools to recruit and retain high-performing teachers**

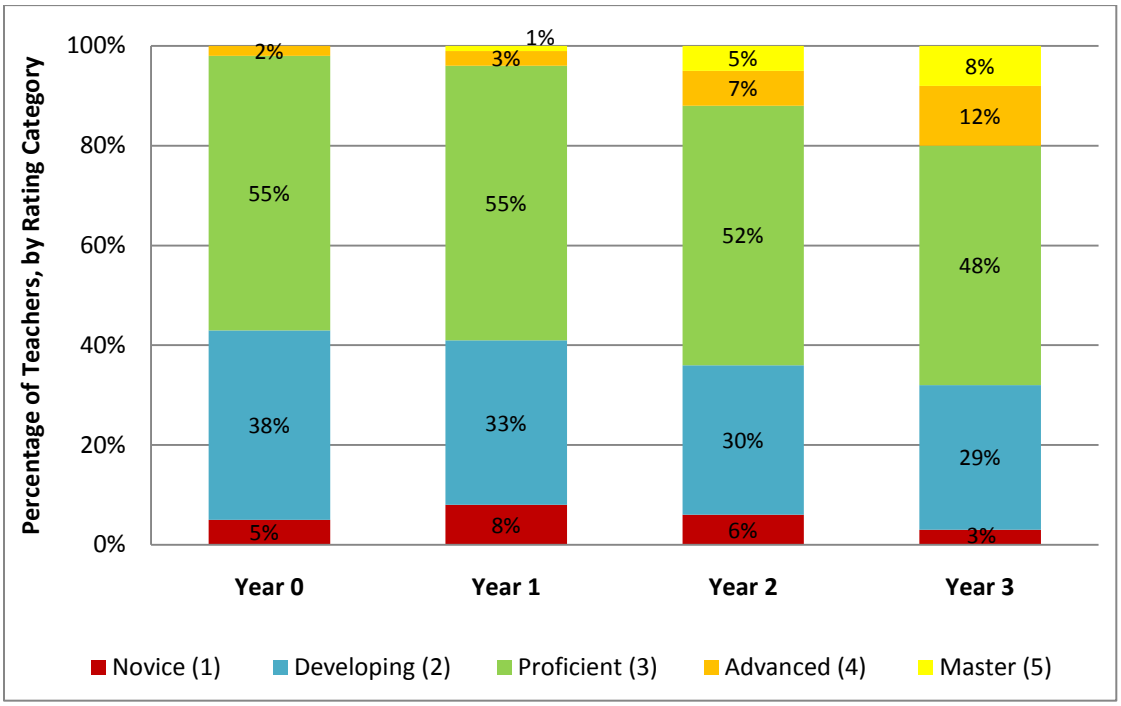
You could track the number of yearly vacancies at the participating high-need schools and the proportion of these vacancies that are filled with teachers qualifying for the signing bonus. You may also want to track how close to the start of the next school year these vacancies are opened and filled, because this may influence schools' ability to recruit high-performing teachers. As noted, tracking the mobility of both signing-bonus hires and existing teachers at the participating high-need schools can help assess how long bonus hires continue working in the high-need schools and whether other teachers—especially other high performers—become more likely to leave these schools.

- **Changes in teacher effectiveness in the high-need schools**

You will likely want to continue using a teacher evaluation rubric—such as the one used to identify teachers who qualify for the bonus program—and apply it to both new bonus hires and other teachers in the participating high-need schools. This rubric could help you uncover the adjustments that new hires experience while teaching at the high-need schools and identify areas where they need support. You could also examine the extent to which other teachers adopt preferred instructional practices, which may help improve

the overall effectiveness of instruction in the high-need schools over time. To examine these changes graphically, you could track changes in the distribution of teacher ratings—based on your evaluation rubric—in the targeted high-need schools. Figure 7 shows an example of a bar graph displaying changes in the distribution of teacher effectiveness ratings, from the year before the bonus program is introduced (Year 0) through the third year of the program (Year 3).

Figure 7: Change in Teacher Effectiveness Ratings in Targeted High-Need Schools



This analysis assumes that the state has a teacher evaluation system that is implemented uniformly and allows comparisons of teacher ratings across schools in the state. Alternatively, you could provide guidelines for teacher evaluation to districts for them to implement individually. In the latter case, teacher ratings may not be comparable across all targeted schools and comparisons would be possible only for schools with a common evaluation system, such as those within a given district. Changes to the evaluation system, or to components of it, would also invalidate comparisons over time.

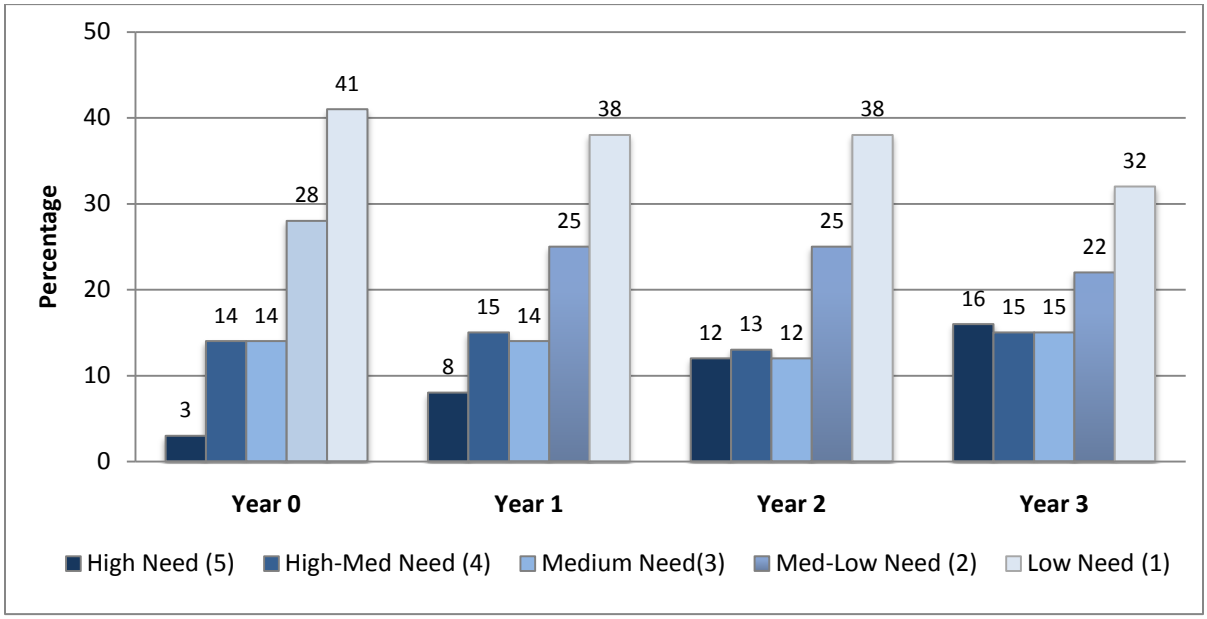
- **Changes in student outcomes in the high-need schools**

Ultimately, you will be most interested in changes in student outcomes. As Figures 5 and 6 suggest, you will want to track not only changes in academic achievement, such as average test scores and proficiency rates, but also changes in student behavior, such as suspensions and other disciplinary actions, and other measures of academic engagement, such as attendance and grade promotion. All may be important intermediate benchmarks of progress.

- **Changes in the distribution of effective teachers across the state**

To determine whether your state is moving toward providing more equitable access to its most effective teachers, you could track changes in the distribution of the highest-ranked teachers across schools over time. Figure 8 illustrates one possible approach. The figure tracks changes in the distribution of the state's highest-performing teachers across schools with various levels of need, from the year before the bonus program begins through its third year.

Figure 8: Distribution of Highest-Ranked Teachers, by Level of School Need



The first step to conducting such an analysis would be to rank schools based on their level of need. To measure need, you could use the same measure used to identify the schools that qualify for the bonus program. Figure 8 shows groups of schools ranked from highest to lowest need. Each group includes one-fifth of the state's schools of a particular type, for example, elementary schools.

Next, using the index that your state uses to determine which teachers qualify for a signing bonus—perhaps teachers who get the top ranking on your evaluation system—you could examine the distribution of your highest-ranked teachers across schools according to need. You could track changes in this distribution of top-tier teachers over time. Figure 8, for example, shows that in Year 0, before the bonus program was launched, only 3 percent of the state's top-tier teachers taught in the highest-need schools, while 41 percent taught in the lowest-need schools. By Year 3 of the bonus program, 16 percent of top-tier teachers taught in the highest-need schools.

To interpret your results, you will need to decide what your state considers to be an **equitable distribution** of its highest-ranked, most-effective teachers. Two ways to define equity follow:

- **Horizontal equity** gives all schools (and their students) equal access to effective teachers; that is, all "equals" receive equal treatment. For horizontal equity, you

would want to see all the vertical bars in Figure 8 reach about equal height, so that schools of different levels of need have about the same share of your most-effective teachers.

- **Vertical equity** distributes resources according to need. Under vertical equity, you would want to see greater numbers of effective teachers working in those schools with the greatest numbers of struggling students. For vertical equity, you would want the distribution of effective teachers in Figure 8 to form a downward slope from the highest- to the lowest-need schools.

In Figure 8, the proportions of highest-ranked teachers across school groupings in a given year always sum to 100 because it is tracking how the highest-ranked teachers in the state as a whole are distributed across schools. Again, this analysis assumes that there is a uniform and stable evaluation system in place that makes possible comparisons of teacher rankings across schools throughout the state and over time. Your state may or may not be developing such a system. Alternatively, you could focus on the proportion of teachers in high-need schools who earn the highest ranking in the state's evaluation system, such as the master teacher category (5) in Figure 7. Your aim would be to make this proportion as large as possible for the high-need schools. This approach promotes vertical equity by working to ensure that as many students as possible attending high-need schools are taught by a highly effective teacher.

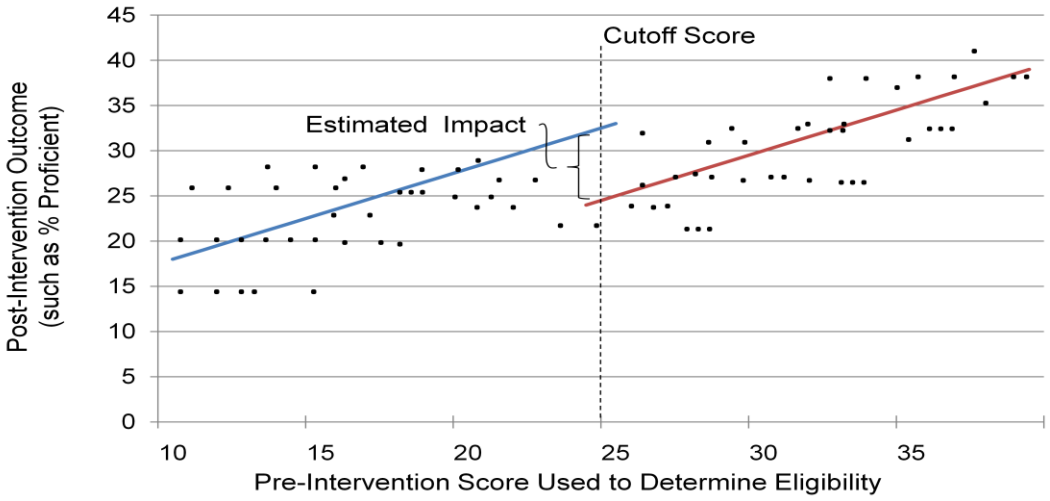
4. How might the state determine whether the bonus program was effective?

Although tracking changes in the distribution of effective teachers and changes in student outcomes will be important and informative, it will not tell you whether it was the signing-bonus program that made a difference. As already noted, for a true assessment of program effectiveness, you need to be able to compare the targeted schools with other schools that are very similar but did not implement the signing-bonus program. This comparison will allow you to determine whether the schools with the bonus program were better able to attract effective teachers and improve student outcomes than they would have in the absence of the program. In the context of this example, factors other than compensation may deter effective teachers from joining the targeted high-need schools.

Equivalent groups of program and comparison schools can be established in different ways. Below are some of these approaches and the circumstances in which they are feasible:

- ***Use the cutoff in ratings of school need to identify a comparison group.*** In this example, the signing-bonus program is targeted at high-need schools and the state plans on using an index of need to determine which schools are eligible for the program. This process creates almost ideal conditions for using a regression discontinuity evaluation design. As long as the state uses a predetermined cutoff point to identify the schools that are eligible for the signing-bonus program (for example, schools with average proficiency rates below 25 percent) and applies this decision rule *without exceptions,* then schools just above this program cutoff point can serve as a comparison group for the schools that qualify for the bonus program. Figure 9 provides a graphical illustration of the regression discontinuity evaluation design.

Figure 9: Sample Regression Discontinuity Results



Note: This figure uses hypothetical data and regression lines to demonstrate a regression discontinuity design with an outcome such as proficiency rates on the y-axis and the score used to determine eligibility for the intervention on the x-axis.

A regression discontinuity design requires a relatively large number of schools right around the eligibility cutoff in order to yield a reliable estimate of the impact of the bonus program. Also, the design is reliable only if the eligibility rule is applied without exceptions. Your evaluation team can provide guidance on these issues.

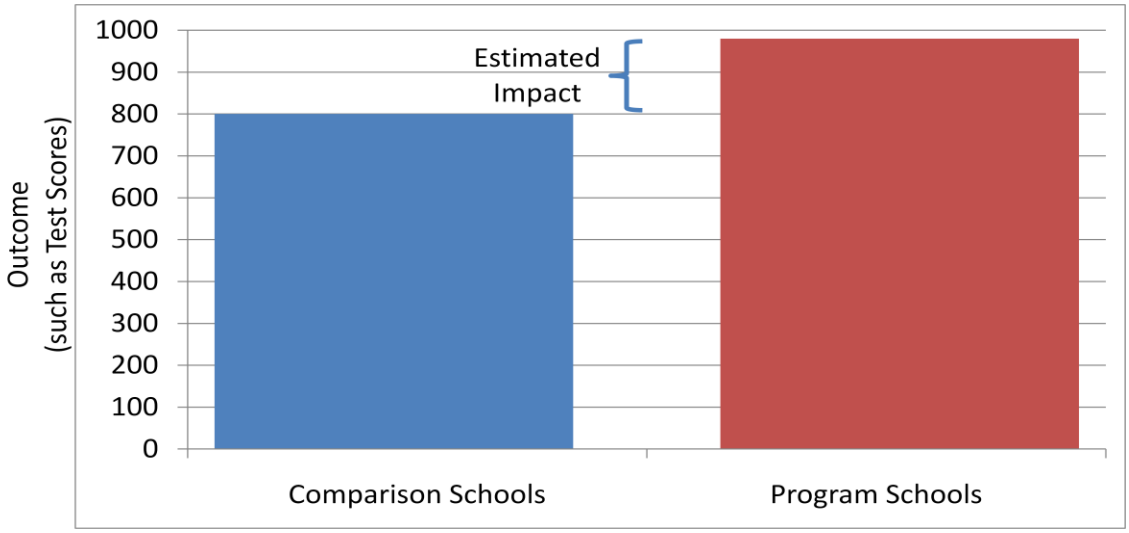
- ***Use a lottery to create equivalent groups of schools.*** In the example, funding for the signing-bonus program is capped at $10 million. If your state has many high-need schools, the share of the signing-bonus funds that any one school receives could be quite limited. **Given this scenario, you might decide that it would be better to pilot the program with a limited number of high-need schools instead.** This would allow you to gather credible evidence about the program's effectiveness before investing more resources and/or rolling it out statewide. Under this scenario, you could pilot the signing-bonus program as a supplement to other teacher recruitment reforms designed to help all high-need schools get priority access to, better develop, or retain effective teachers.

  To implement this design, you would first identify all eligible high-need schools and then use a lottery or another random assignment process to decide which schools receive access to signing-bonus funds as part of the pilot phase. As discussed in Chapter II, this could be as simple as listing all eligible schools in an Excel spreadsheet and using the program's random-number generator to assign a number between 0 and 1 to each school. Then, you would select the top or bottom "X" schools for the bonus program, depending on the number of schools that you want to participate in the bonus program. The decision rule to be applied must be set before assigning the random numbers and no numbers should be reassigned.

  By using a random process to select the signing-bonus schools from among equally needy schools, you would create two groups of schools (bonus and comparison) that would be statistically equivalent. You would then be able to compare key outcomes

for these two groups of schools over time to measure the impacts of the bonus program. An attractive feature of this study design is its relative simplicity when it comes time to estimate the effects of the bonus program on key outcomes. You would just compare the (signing bonus) schools and the comparison (no bonus) schools to estimate the impact of the signing-bonus program on the outcome(s) of interest. Figure 10 shows how findings from a random assignment evaluation of the bonus program might be presented.

Figure 10: Sample Findings from a Random Assignment Study



Note: This figure uses hypothetical data to demonstrate a random assignment design with an outcome such as student test scores on the y-axis. The difference between the average outcome for members of the "program" group and the average outcome for members of the "comparison" group is the estimated impact.

- ***Use a lottery to compare different versions of the bonus program.*** A bonus program like the one in this example could provide other important types of information. For example, you might want to find out what bonus amount is sufficient to attract effective teachers to high-need schools or what payment structure works best. Is offering a smaller-but-full bonus up front more effective than offering a larger bonus that is paid in installments? By designing different options and using a lottery to assign them to schools, you could test the relative effectiveness of different types of bonus packages. This approach would not answer questions about whether the bonus program is really worth adopting in the first place. However, it would tell you which bonus scheme is most effective (or cost-effective). For this design to yield reliable answers, a sufficiently large number of schools must be assigned to each bonus scheme being tested. Hence, it might be best to restrict your testing to a few approaches.

Example 2: A Teacher Recruitment, Retention, and Improvement (TRRI) program

A state initiates a two-part program. Compensation strategies to attract effective teachers to work in high-need schools and to retain and reward effective teachers already working in high-need schools are paired with targeted professional development and coaching to raise the capacity of existing teachers to work with high-need students. The teacher compensation strategies are (1) signing bonuses to attract effective teachers from other schools, (2) retention bonuses to discourage attrition or transfers of effective teachers already at the high-need schools, and (3) performance bonuses tied to marked improvements in student achievement and reductions in achievement gaps to reward excellence among all teachers. The PD strategies include (1) identifying areas of needed support and growth for teachers at the high-need schools and (2) providing targeted PD and coaching to increase their capacity and effectiveness.

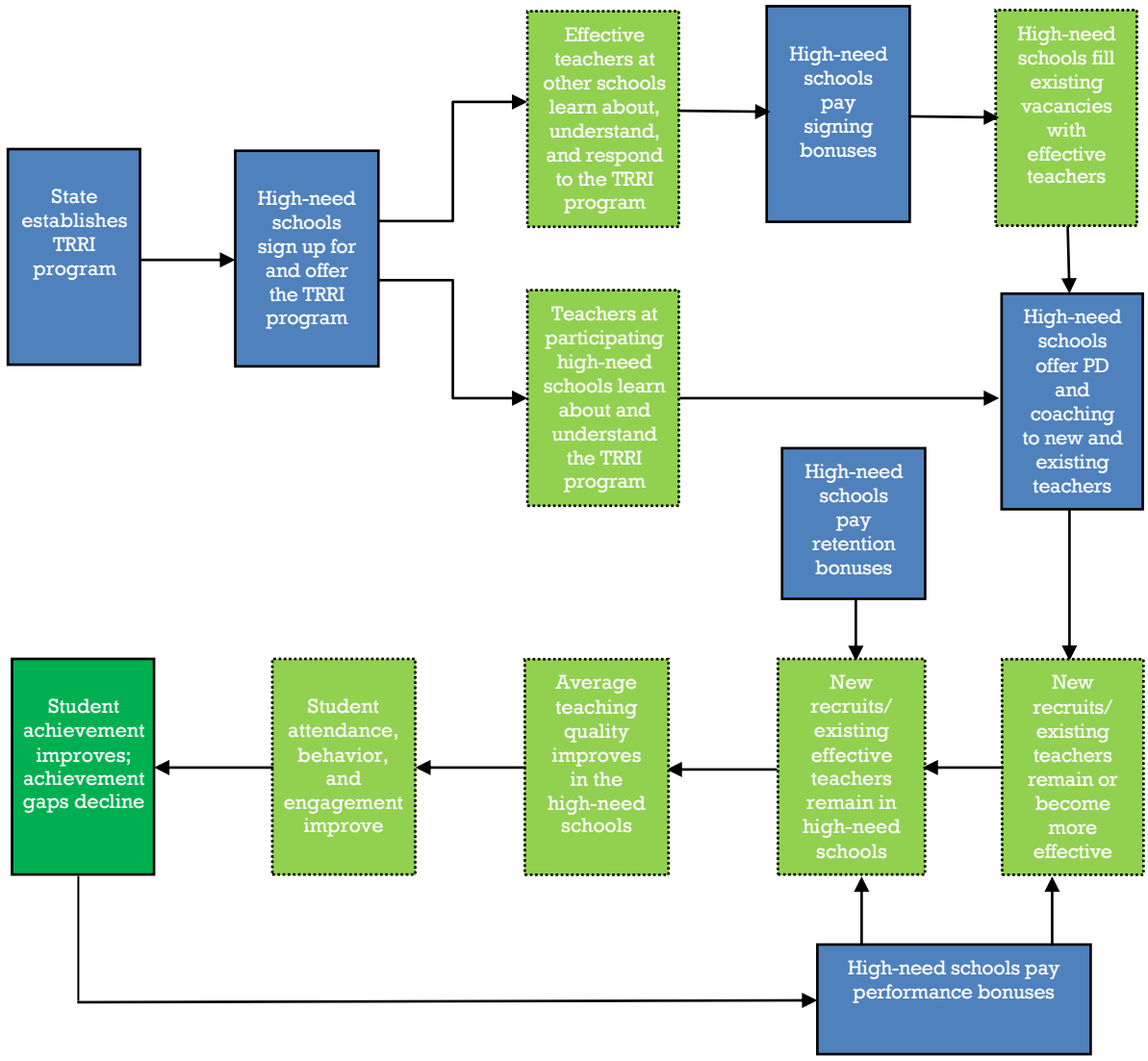
Compared with the first example in this chapter, this example is noticeably more complex. Not only does it involve multiple teacher compensation strategies, it also combines them with other PD and coaching efforts to strengthen teachers' skills. Evaluating comprehensive reforms such as this one is not simple, but the same general principles for evaluating more narrowly defined reforms apply. Much can be learned from evaluating these complex, comprehensive reform efforts. In fact, as states increasingly invest in comprehensive reform initiatives, it becomes imperative to examine the results of these efforts to extract information to guide future initiatives. Below are some issues you might consider if your state were adopting the TRRI initiative described above, or a similar multicomponent reform program.

1. What is the state trying to accomplish through this initiative?

As in the previous example, the state is ultimately interested in accelerating learning and increasing achievement among students in high-need schools and reducing gaps between these students and other students throughout the state. However, much must happen between rolling out the TRRI program and reaching that ultimate goal. As in the previous example, it would be helpful to map out the chain of inputs or actions and outputs or results that you hypothesize will lead to the desired changes in student learning (Figure 11).

Although the TRRI program is more complex, the change model for this program shares many elements with the change model for the signing-bonus example (Figure 6).

Figure 11: A Change Model for the TRRI Program



The TRRI change model maps how the individual strategies of this initiative—signing, retention, and performance bonuses; PD and coaching (shown in blue)—are part of a logical sequence of events to take place in the targeted high-need schools. It also includes some of the desired changes, such as interim outcomes (shown in light green), expected to come about as a result of the program in the targeted schools. Although all the teacher compensation and PD strategies are intended to help bring about the same ultimate outcome—improved student achievement and decreased gaps—the pathways by which they contribute to this outcome differ and target different segments of the staff within the high-need schools. A detailed change model such as this can help you identify linkages worth testing and important indicators to track.

2. How will the state know whether the initiative is implemented well?

Conducting an implementation analysis is particularly important for multicomponent programs like the TRRI initiative. When you pursue multiple strategies simultaneously, it can

become quite challenging to keep track of the myriad lessons that emerge as the program's individual components are implemented. Therefore, you will likely want to focus on the most important implementation questions relating to individual strategies and the initiative as a whole.

Issues discussed above--about who qualifies as an effective teacher and which schools qualify as high need--are also relevant for the TRRI program. For the signing bonuses, the issue of publicity to teachers at other schools is likely to be important. By contrast, for the retention bonuses, implementation questions may focus on how well these bonuses are targeted—that is, which teachers at the high-need schools are offered bonuses—and whether the bonus amount represents a sufficiently persuasive incentive to keep these teachers at these schools. For the performance bonuses, implementation questions may focus on how the bonus dollars are allocated among teachers within the school and how the awards influence the behavior of both recipient and nonrecipient teachers. Chapter II discusses some of the implementation questions that are relevant when evaluating such efforts including, for example, the design and intensity of the training being provided.

For the overall TRRI program, you might also want to understand how the component strategies interact with, reinforce, and influence one another. For example, does the influx of new effective teachers from other schools undermine efforts to use PD and coaching to bolster the skills of and motivate existing teachers at the high-need schools? How do existing staff respond to dollar-amount differences between the signing and retention bonuses? Examining these questions can help you make sense of changes in outcomes that you may observe (or fail to observe) later, as well as provide valuable lessons for others adopting similar reforms.

3. How might the state track changes in outcomes?

As the change model (Figure 11) suggests, the TRRI program has a wide range of relevant interim outcomes. As with implementation, you may therefore need to be strategic when deciding which outcomes to track. You might prioritize outcomes based on your relative investment in the components of the program (for example, signing versus performance bonuses). Another important factor might be the costs of data collection. Suggestions for some key outcomes to track are as follows:

- ***Changes in teacher characteristics, instructional practices, and effectiveness.*** You might want to track which teachers respond to the different types of incentives introduced under the TRRI program. Because the program aims to enhance capacity and to reward performance improvements, tracking changes in teachers' instructional practice is also likely to be important. As mentioned in Chapter II, you could develop observation protocols to track changes in instructional practice in areas that are the focus of the PD training. Also, as in the signing-bonus example, you could use a teacher evaluation rubric—like the one used to identify teachers who qualify for the signing bonus—to evaluate both new hires and existing teachers in the targeted high-need schools to track changes in teacher effectiveness over time.

- ***Changes in school climate and academic achievement.*** Again, changes in staff and student attendance, student disciplinary actions, staff turnover and morale, and other measures of academic engagement and school climate may be important precursors to the changes in student achievement that are the ultimate goal of the TRRI program.

As the program's change model suggests, the same student achievement outcomes discussed under the signing-bonus example are relevant for the TRRI program.

- ***Changes in the distribution of effective teachers relative to school need.*** Owing to the greater intensity of the TRRI reforms, you may expect to see larger changes in the distribution of your top-tier teachers over a shorter period of time. If the professional development and performance incentive elements of the initiative prove successful, you would also expect to see growth in the number of teachers evaluated as being highly effective.

4. How might the state determine whether the program was effective?

Throughout, this guide has stressed the importance of identifying a credible comparison group so that you can estimate the effects that you can confidently attribute to the program being evaluated. Evaluating a comprehensive initiative like the TRRI program may seem difficult. Such comprehensive reform initiatives tend to be part of an ambitious agenda to address persistent inequities across students and/or schools. Hence, such initiatives may be rolled out to all eligible schools at once, which affects the kinds of questions that an evaluation can answer.

Nonetheless, an ambitious initiative like the TRRI program can be evaluated effectively. Capacity constraints—such as in the availability of funds for bonuses or of PD trainers or coaches—may offer opportunities to pursue a rigorous evaluation. If the program is rolled out to all or nearly all eligible schools at once, you can still compare alternative versions of it, in the following ways:

- ***Capitalize on capacity constraints: stagger interventions to create comparison groups.*** Implementing a complex bundle of compensation and PD reforms requires more capacity than implementing a single initiative. If capacity constraints in your state or district dictate rolling out the TRRI reforms in stages, they may provide an opportunity for comparison. Using a random process such as a lottery to select schools or districts to receive the TRRI reforms first will enable you to compare outcomes for early adopters and late adopters.

- ***Use a lottery to test different versions of the TRRI program.*** It may be worthwhile to investigate whether some combinations of the TRRI strategies are more effective than others. For example, schools or districts in your state may be most interested in the capacity-building PD and coaching. If so, you could roll out these TRRI elements universally, and select high-need schools by lottery to receive the bonus components. This would present an opportunity to examine, for example, whether performance bonuses offered on a yearly basis coupled with job-embedded PD and coaching are more effective than the PD and coaching alone.

[This page was intentionally left blank for double-sided printing.]

Evaluating Strategies to Turn Around Low-Performing Schools

by Rebecca Herman, Daniel Aladjem, and Kirk Walters

States and districts have been working to turn around their lowest performing schools for many years and have employed multiple strategies to promote school improvement. Under ARRA, states will implement at least one of the four required models (see text box) in schools identified as persistently low performing. Ultimately, all states will face the same challenge of determining the extent to which those schools have improved. This chapter focuses on strategies you may be considering for turning around low-performing schools to illustrate how you can incorporate evaluation into your new or revised programs.

How Turnaround, Restart, Closure, and Transformation Work

- ✓ **Turnaround:** The principal and at least half the staff are replaced, and the instructional program is revised. Turnaround is designed to bring in new, highly qualified staff and new programs, training, and support, often through a packaged reform model.
- ✓ **Restart:** The school is closed and then reopened under the direction of a charter or and education management organization (EMO). Restart presumes that private operators will foster greater innovation and improvement than public school districts.
- ✓ **Closure:** Schools are closed and the students attend other schools in the district. Closure eliminates schools that are considered beyond repair and is intended to offer students a better chance for success at another school.
- ✓ **Transformation:** Changes are made in professional development, instruction, curriculum, learning time, and operating flexibility. Transformation assumes that the core instructional staff at a failing school are competent but need new leadership, programs, training, and support, often through a packaged reform model.

As discussed in the Introduction, it is best to think about evaluation when you are developing your plans for a new program. This way, you can be sure to have the information you need to answer questions later. It also can be easier and more efficient to build in data collection from the beginning. Two examples illustrate how the evaluation framework in the Introduction might apply to a turnaround initiative that a state might actually implement. The first example looks at the overall impact of a state's use of charter management organizations (the "restart" option). The second example considers the use of whole-school reform models and how states may want to monitor implementation to support positive outcomes (one possible way of approaching the "transformation" option).

There are two critical differences between the two examples. First, the nature of the intervention differs. Example 1 involves opening a school with new staff and management, wheareas Example 2 involves changing the operations within an existing school. Second, the nature of the evaluation question appropriate to each differs. Example 1 focuses primarily on *whether* schools have improved, and Example 2 looks also at *how* they improved. These examples were deliberately written to provide different ideas for evaluation. In your state, you may end up with some variation or hybrid of these examples.

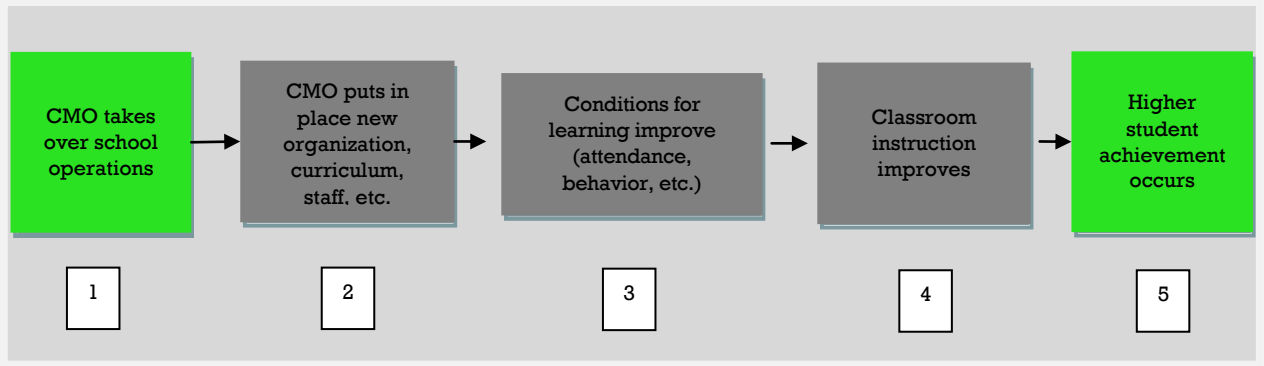
Example 1: Assessing the Overall Impact of a Restart Strategy

A state has identified 60 low-performing schools. Thirty of these schools will be closed and reopened as charter schools under the direction of charter management organizations (CMOs). The other 30 schools will do some mix of reforms negotiated with their districts. This state is interested in examining the extent to which, at the end of four years, this strategy turned around these 30 schools.

1. What is the state trying to accomplish through this turnaround strategy?

The goal of this strategy is to improve student achievement by contracting out the design and management of low-performing schools to charter management organizations. These CMOs will be independent of bureaucratic control and be more innovative and responsive to student needs. Although many changes are associated with a restart, such as changes in curriculum, staff, school organization, and others, the state is focused on whether restarting the schools under CMOs improved achievement compared with schools that did not restart as CMOs but used other strategies for school improvement. In this example, the state is not focused on learning about the specific reform strategies used by the CMOs.[3] Figure 13 shows the change-model through which the CMO restart aims to improve achievement.

Figure 12: CMO Restart Change-Model



Using Figure 13 as a framework for an evaluation, Step 2 focuses on implementation—did the CMO put in place the changes it had planned? Steps 3 and 4 show intermediate outcomes—did the reform change important aspects of the school climate and operations? Steps 2 to 4 aren't the primary focus for this example but are explored in greater depth as part of Example 2. Step 5—student achievement improvement—is the focus of this evaluation example.

[3] In many ways, the ideas discussed in this example are also relevant for evaluating transformation or turnaround approaches, when the interest is primarily about whether the turnaround model improves student achievement outcomes.

2. How will the state know whether the strategy is implemented well?

For this example, the primary evaluation goal is to assess whether restarting under a CMO allows the persistently low-performing schools to improve student achievement. However, the state should still keep track of some basic benchmarks to determine whether the intervention actually happened. For instance, if a school failed to open on time or the CMO was never able to get its program up and running, the state would want to know. The text box shows some simple pieces of information that the state may track on its own or request from the operating CMOs.[4]

Tracking Basic Implementation for CMO Restart

- ✓ Did the school or schools open on time?
- ✓ Were key staffing positions filled with permanent staff?
- ✓ Were adequate numbers of students enrolled?

3. How might the state track changes in outcomes?

The primary question for this example is, Has the CMO restart strategy improved achievement in the targeted schools? Generally, states already track student outcomes, but you can now see whether scores for students in targeted schools appear to improve after the initiative has been rolled out. It is particularly important to consider these gains in relation to years past. If you have multiple years of prior achievement data available, you can compare achievement **trends** before and after the implementation of the turnaround strategy by using these longitudinal data for the targeted schools. To better understand where and how the trends might be changing, you could look at achievement by grade level and subject area at participating schools. Of course, to truly know that the school has improved (and not just that the mix of students has changed), it is essential to do an analysis at the student level (see text box on following page). In fact, it is very likely that the student composition will change in the schools in this example. That change might completely explain any improvements in outcomes (see Question 4 for more discussion on this issue).

To know whether the school turned around, you will need to establish a threshold for what you consider "enough" improvement. For example, you might explore whether the gains were substantively important, such as students, on average, gaining the equivalent of two grade levels in one year, or statistically significant, such as the proportion of students meeting proficiency standards increasing significantly. You might also look at whether the scores at the end of a given year were acceptable relative to a set goal, such as at least 75 percent of students scoring proficient. For issues and resources related to analyzing student achievement data, see the text box below:

[4] This list might look a little different if you are looking at turnaround, closure, or transformation approaches instead of restart. For example, you might look more closely at professional development and less at student enrollment for the transformation approach.

Challenges in Measuring Turnaround Success

Resolve issues related to analyzing achievement data:

- ✓ ***Should analyses be conducted at the student or school level?*** It is important to use student-level data, which makes it possible to account for students who change schools. If analyses are done at the school level (looking only at the school average achievement), you may not know whether gains or losses are due to changes in student composition, with the better students leaving (or coming to) the school.
- ✓ ***Will analyses use scale scores or percent proficient (for example, 34 percent of students did well enough on the state test to be rated "proficient" or better)?*** If at all possible, use scale scores. Using percent proficient as the outcome is problematic because students who move from just below the cutoff to just above can make gains look better than they really are (the "bubble effect"). (See, for example, Ho, 2008.) If it is necessary to use percent proficient, how will your analysis address the bubble effect?
- ✓ ***How do you look at gains when a state doesn't have vertically equated tests?*** For example, what growth analysis can you do when scores in grade 5 aren't on the same scale as scores in grade 4? (See, for example, Patz, 2007.) At this point, most states are (or should be) ensuring that their tests are vertically equated.
- ✓ ***How will analyses account for the fact that students in the same class and students in the same school will have some common experiences that make the observations not entirely independent?*** The analysis should take into account that students are "nested" within classes within grades within schools. Students in the same class (or school) may be similar in ways that have nothing to do with the CMO restart (for example, they all had doughnuts just before taking the state test and were too jumpy to focus) (See Raudenbush et al., 2004.) The sample and the analysis can be set up to account for such similarities within classes or schools.
- ✓ ***Will the analyses look at changes from one cohort to the next (say, grade 4 in 2010 and grade 4 in 2011) or growth of individual cohorts (say, grade4 in 2010 and grade5 in 2011)?*** What are the tradeoffs, such as bias and cost, for each approach? (See Choi, 2009; Feldon & McKinlay, 1993; and Goldschmidt, Choi, Martinez, & Novak, under review.).

Resources to resolve these issues:

- ✓ ***Expert consultation.*** Search on the What Works Clearinghouse Evaluator Registry or through other resources. Select experts based on expertise in your state assessment and standards, adequate yearly progress (AYP) and other national guidelines, complex achievement analyses, and school turnaround.
- ✓ ***Recent research.***
  - National Council on Measurement in Education
  - Turning Around Chronically Low Performing Schools Practice Guide (see especially the definition for turnaround schools, pages 4-5)

4. How might the state determine whether the turnaround strategy is effective?

Just knowing whether outcomes have improved does not tell you that it was this strategy that made a difference. To truly assess program effectiveness, you need a rigorous design. (See the guide's Introduction for a more general discussion of strong evaluation designs that can help you draw conclusions about whether a program has worked.) Some of the challenges to doing a strong study of school turnaround—especially for restart schools—are (1) focusing on the same population of students throughout the study, (2) finding good comparison groups and ensuring nonbiased assignment to groups, and (3) having enough schools to study and variations in the reform being studied. These issues are discussed below:

- **Focusing on the same student population.** In low-performing schools, student turnover is often high. In restart schools, especially, it is very likely that the students who enroll in the restart school will be different from the students in the former school in important ways. For example, the new students might be higher achieving or more motivated than the former students—they and their parents might seek out the restart school as a good opportunity.

  This difference could present a policy problem and an evaluation problem. From the policy perspective, if stronger students attend restart schools and the most at-risk students relocate to other, low-performing schools, the most at-risk students are not benefitting from the restart. From an evaluation perspective, gains in achievement from having different, higher-achieving students might be mistakenly attributed to the restart strategy.

  Therefore, in a restart evaluation, it is important to figure out whether students after the restart are substantially different from students before the restart. If they are, an evaluation of the school will not be able to separate the effects of the student differences from the effects of the restart strategy. The evaluation would need to be reframed from "Did restart improve learning in the school?" to "Did students who were in the school before restart benefit from the restart?" The analysis would need to be at the student level and would look at the gains of students who were in former school *regardless of which school they ended up in*.

  If the students after the restart are similar to the students before the restart, it is possible to conduct an evaluation as described below.

- **Finding the right comparison group.** To truly assess effectiveness, you need a fair and realistic group of nonparticipating schools with which you can compare the restart schools. This comparison helps you determine whether students in the restart schools do better than they would have done without that support. This analysis can be particularly hard to do for a turnaround study because there is likely to be a reason some schools were selected for restart—particularly low scores, poor school management, unusually long history of low achievement—that would make those schools different from other persistently low performing schools even before the intervention. Further, one of the greatest challenges in evaluating the effectiveness of programs for turning around persistently low-performing schools is that it is very difficult to find persistently low-performing schools that are not doing *some* kind of reform. Most underperforming schools are under substantial pressure to do *something*. So, to describe the findings accurately as "CMO restart versus other interventions, "it is extremely important to ensure that the comparison is not doing the essential elements of the restart. For example, the comparison schools were not operated by a CMO or closed and reopened.

  Randomized control trial: The most effective study design for an evaluation study is a randomized control trial. In this design, a sample of schools is identified and then schools in the sample are assigned by lottery either to participate in CMO restart or to be in the comparison group. In this example, the state might have identified 60 low-performing schools but has resources to monitor only 30 restart schools. The state might randomly select 30 schools for the restart option. The comparison group might do business as usual—whatever mix of reforms these schools would do if they hadn't been selected for this study. This design would be appropriate if your study

was answering the question, Does CMO restart have a greater impact than a mix of other approaches? Or, the comparison group might be assigned to a particular CMO for restart. This design would be appropriate if your study was answering the question, Does one CMO restart work better than another?

How the State's Turnaround Approach Influences Evaluation Options

IF many schools are involved in turnaround AND the state has a clear, prescriptive approach, a rigorous quantitative evaluation is possible.

IF few schools are involved in turnaround and the state allows schools to vary in their approaches, a qualitative study is likely the best option.

However, for political or logistical reasons, it's often difficult for states to use a lottery to assign schools to CMO restart versus other turnaround reforms. Alternative research designs that are particularly appropriate for this example are comparative interrupted time series, regression discontinuity, and matched comparisons.[5]

Regression discontinuity: For this design, you would separate your persistently low-performing schools into two groups based on whether they are above or below a pre-established cutpoint on an important variable. For example, you could rank-order the 60 low-performing schools according to the average percentage of students proficient in reading and math (either one or both) across tested grades for the last five years. Then you might assign the lowest 30 schools to the CMO restart and the other schools to the comparison. If the intervention is effective, you will see more improvement in CMO restart schools than in comparison schools, even accounting for initial differences between the two groups. It is important to have a clear cutpoint that separates the two groups and to strictly assign schools to groups based only on whether they are above or below that cutpoint. This type of research design is reliable only if there is no fuzziness about the cutpoint. If a state lets a school squeak by to be in the restart or comparison group, the design becomes less reliable. This design also needs a large number of schools, perhaps more than the 60 available in this example.

Comparative interrupted time series: For this design, you would compare the patterns of achievement in the CMO restart schools and those in other comparable schools before and after the CMO restart, using data for multiple years. If the intervention is effective, you would see more improvement in the CMO restart schools compared with other schools after the turnaround initiative was introduced. For this design, it is important to (1) compare the achievement patterns in the CMO restart schools with patterns for the same years in similar schools not subject to CMO restart; (2) have good, comparable data for several years before the CMO restart; and (3) have a clear point at which the intervention started (to help separate out before from after). This design is manageable if the state has good data for several years before the intervention is introduced.

[5] These research designs also would work for the turnaround and transformation approaches. However, the closure approach would require an entirely different approach, such as comparing the achievement trajectories of students in one school (before it closed) with the same students' trajectories in their new schools and determining whether students in the closed schools grew more than similar students who did not attend the closed schools.

Matched comparison: Both of the study designs above have comparison groups. Although not as strong as a randomized study, they are particularly strong designs among nonrandomized studies. A somewhat weaker design is a simpler comparison study in which outcomes for the CMO restart schools are compared with outcomes for similar schools. The findings from such an evaluation are less reliable because only one pretest is used to identify comparable schools, instead of using multiple data points before and after the intervention (interrupted time series) or fully modeling for differences between the groups (regression discontinuity).

The discussion up until now has focused on designing a study with a good comparison group. One pitfall, particularly for restart schools, is that no matter how good the design, changes during the intervention (and study) can affect the comparison. Because the school is closing and reopening, there is a chance that different students will attend the restarted school than had attended the closed school. If that were to happen, you would not know the extent to which improved achievement reflects differences between the old and new groups of students or the reforms that came with the restart. For a good evaluation (and also to make sure the students in the old school are getting access to the new and—potentially—better school), serving the same (or very similar) students is helpful. When this isn't possible, it is important to examine the extent of changes in the student population served by the restart schools and include this information in the description of outcomes of the reforms. If possible, the descriptions should match students based on their background characteristics.

- **Sample size and reform variation.** Another challenge in studying turnaround schools is sample size, that is, the number of schools in the evaluation. Although there are many persistently underperforming schools that could use support, Federal education initiatives suggest focusing intensively on a limited number at first, such as the lowest 5 percent. Sample size needs to balance practical needs with requirements for having useful information from the evaluation. Having fewer schools helps the state target its resources. However, too few schools can mean that the evaluation does not have sufficient ability to identify the real effects of the restart approach (called statistical power).

  This balance is especially true for a reform like restart, which can have substantial variation across schools in how the intervention is implemented. For example, one CMO might departmentalize an elementary school (with students having different teachers for different subjects), and another CMO might keep students in the same class all day. Or, each CMO might choose a different curriculum and instructional strategies. Ideally, each CMO will submit a plan to the state describing intended implementation. With only a few schools in the evaluation, these differences can make it difficult to separate the specific CMO approach from the effects of CMO restart overall. These school-to-school differences become less critical with larger samples. If the state has very few schools identified for restart CMO, one option is for the state to limit the number of CMOs to be evaluated and choose these carefully, based on clearly distinct philosophies.

  Overall, the state's approach to school turnaround makes a difference in the type of evaluation that is possible. The number of schools and the uniformity of the reform (which depends on how prescriptive the state is) affect evaluation designs.

Example 2: Whole-School Reform as a Transformation Strategy for Low-Performing Schools

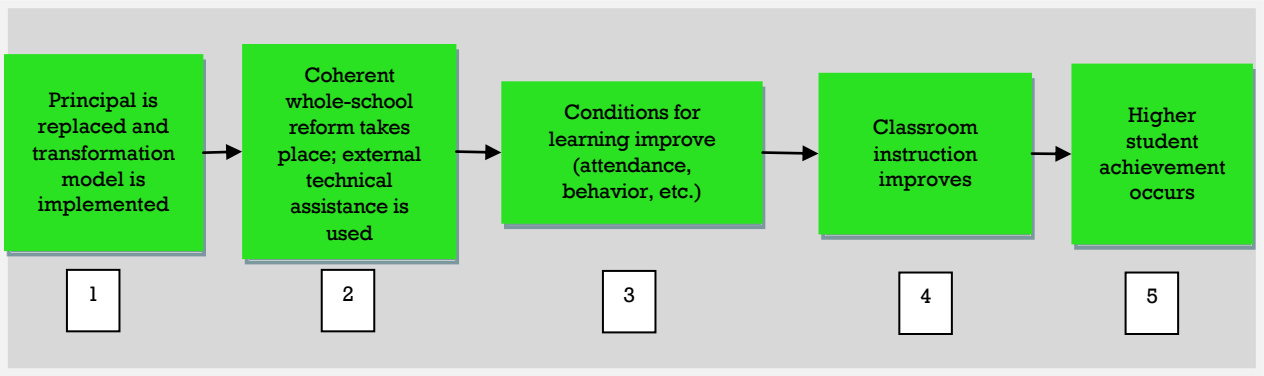
A state has contracted with a national, whole-school reform developer to implement its program under the transformation option in the state's low-performing schools. This program has been validated through research and specifies curricula, assessment, and professional development elements to be implemented as part of its model. The state wants to both assess the impact of the whole-school reform on student achievement and understand exactly how the selected reform approach influences student achievement.

1. What is the state trying to accomplish through this initiative?

The state is adopting a whole-school reform model to improve some of its lowest-performing schools. Unlike the restart example, the state aims to improve schools by using primarily existing schools, staff, and students. As in all school reform efforts, the state is most interested in whether the reform substantially improved student learning.

However, intermediate outcomes are also crucial to the state's initiatives. School reform research indicates that schools can take three or more years to achieve significant gains, with the rare exception of a few turnaround schools that make major gains in one or two years (Desimone, 2002). Intermediate outcomes can signal whether the reform effort is on the right track—that changes likely to improve student learning are happening—so the school can continue its positive direction. Without these early signs of success, it is harder for a school to know that it is doing the right things and should stay the course. Similarly, these early signs of progress can convince the community to continue to support the school. For example, signs of improvement might convince parents to keep their children in the school another year. Finally, intermediate outcomes that are later linked to achievement gains can give confidence to replicating schools that they are pursuing a promising approach. Figure 13 shows the change-model with which the whole-school reform aims to improve achievement.

Figure 13: Whole-School Change-Model



In the whole-school change-model, Steps 1 and 2 focus on implementation. Was the principal replaced? Were the components of the whole-school reform model put into place? Steps 3 and 4 show intermediate outcomes—did the reform change important conditions and/or processes in the school? These changes should, in turn, improve student achievement (Step 5).

2. How will the state know whether the initiative is implemented well?

The first question is whether the core elements of the transformation strategy were put into place as planned (Steps 1 and 2 in the example above). It is therefore important to monitor implementation so that you will know what turnaround actions were actually taken and how well they were done. More specifically, monitoring implementation will show (1) where schools need more support to successfully implement the reform components and (2) how implementation relates to student achievement outcomes.

Monitoring implementation does not need to be costly and time-consuming. The model developer will likely have detailed implementation benchmarks and rubrics to assess how well the schools are implementing all aspects of the model. Developers often include the service of tracking implementation as part of the package, which reduces the burden on the state. At minimum, however, the developer should provide implementation benchmarks or rubrics and training in how to use them. In addition to these implementation benchmarks or rubrics, states can also consider developing their own implementation checks, using information from research on school reform. For more information on the implementation of whole school reforms, especially implementation indices, see Aladjem and Borman (2006).

One of the greatest challenges in implementing turnaround, whether using whole-school reform or another strategy, is rallying support for the effort. Staff, students, parents, and others involved with persistently low-performing schools have usually witnessed years of failed improvement efforts, especially when the school has tried to improve using existing staff. Case study research on turnaround schools shows that most schools that successfully turn around use some kind of "quick win" early in the turnaround process. Quick wins can help communicate to school staff and the community that it is possible to make a meaningful improvement quickly—and worth the time. For example, the principal might have the school landscaped or painted before the start of the school year or change the schedule to establish dedicated core instructional time. Quick wins can help rally staff and others to put real energy into implementing the reform. For a reform effort like whole-school reform that uses existing staff, a state may want to look at whether there was a quick win and whether it had the desired effect of motivating staff. (See the text box for some ideas; see Doing What Works for materials on quick wins.)

Quick Wins

Did the school have a quick win?

- ✓ Did something happen to **markedly** improve the daily lives of teachers and students?
- ✓ Did it happen within a month of starting the effort?
- ✓ Was it accomplished with almost no additional money or authority?

Did the quick win support implementation?

- ✓ Minimal: Did staff and students notice the change?
- ✓ Moderate: Did more staff indicate that they would support the reform (especially those who had been on the fence) after the quick win?
- ✓ Substantial: Did more staff invest more time and energy into the reform after the quick win?

You can use interviews and climate surveys before and after the reform to look at the changes in attitudes.

3. How might the state track changes in outcomes?

- ***Intermediate outcomes.*** As noted above in Figure 13, intermediate outcomes can include better conditions for learning and better classroom instruction. Some indicators of school conditions for learning, such as attendance and discipline, are

easy to measure and are good early signs that the school is improving. Further, they are often necessary conditions for improving instruction. Improved instruction is harder to measure but is extraordinarily important. Educational research and theory suggest that improved instruction has the greatest direct impact on student achievement. See Table 3 for just a sample of the many possible intermediate outcomes that could be explored.

Table 3: Intermediate Outcomes

Possible Outcomes	Measurement Strategies
Conditions for Learning	
Attendance: Are more students in school for more of the day?	School and district records. No additional data collection is needed.
Discipline: Are there fewer and less severe disciplinary issues, reducing distractions from core academic mission?	School and district records. No additional data collection is needed.
Staff: Do teachers have relevant knowledge to teach their subject? Are the best teachers staying with the school? Is a strong turnaround leader in place?	School and district personnel records. Consider a survey of teacher knowledge. See Doing What Works for teacher knowledge and skills inventory.
Parent, community involvement: Are parents and the community more involved in volunteering or otherwise supporting the school? Are parents more involved with their children's schoolwork?	Surveys or focus groups. Many schools and PTOs already collect these data; consider whether existing relevant data exist.
Additional support for students: Does the school offer research-proven supports for at-risk students struggling with academic, social emotional, or health issues? Are the students who most need these supports getting them?	Interview with principal or other core school staff, student participation records.
Climate: Does the school community (staff, students, parents) agree on school goals? Is the focus of the school on academics? Are distractions from core focus reduced?	Climate survey. Many states and districts conduct climate surveys, sometimes to meet Federal grant requirements; consider whether existing survey data or an expanded version of a current survey could be used (see, for example, CASEL's list of climate surveys).
Data use: Does the school have access to a system for collecting, analyzing, and using data to improve learning? Are data disaggregated at the student, class, and school levels? Are teachers trained to use data and held accountable for doing so?	School and district interviews, document review, survey. Sometimes the approach to data use is district-wide; consider whether the district already collects data on its approach and progress (see also the Survey of Education Data System and Decision Making).
Instructional Improvement	
Curriculum: Is a research-based curriculum in place for core subjects? Is the curriculum aligned to state standards and assessments?	District interviews and records, observation (see Doing What Works tools on improving instruction).
Instructional time: How much time is spent on instruction (before or after school, intervention period during the day)? Is core instructional time protected from interruptions?	Review of school documents (such as class and year schedules) to determine total amount of instructional time, combined with observations of instructional time and time on task to determine academic focus of available time.
Instruction: Is instruction of high quality? Are effective practices (differentiated instruction, formative assessment) being used? Is time-on-task high?	Classroom observation ofcore classes. See Doing What Works tools on improving instruction.

- ***Student achievement***. Many of the points made in Example 1 about measuring student achievement apply to this example as well. In addition, the state might consider collecting and analyzing student achievement data more frequently to better determine the linkage between stages of implementation and achievement gains.

4. How might the state determine whether the turnaround initiative was effective?

The research designs explored in Example 1—randomized control trial, comparative interrupted time series, regression discontinuity, matched comparison group—are also relevant for this example. Because there are so many different approaches to whole-school reform, it might be possible to design a rigorous evaluation where schools are assigned to different whole-school reforms by lottery. This design has two unique features: (1) random assignment, which is the strongest research design for outcome evaluations, and (2) head-to-head comparisons, in which you compare different interventions (in similar schools) to see which works best. One advantage of a head-to-head comparison is that all schools in the sample get something. However, that does change the nature of the study from Does whole-school reform turn schools around? to Which whole-school reform works better to turn schools around?

If using a head-to-head comparison, it is helpful to compare *very* different types of whole-school reforms. For example, a head-to-head study might contrast a highly structured, prescriptive reform with an approach that is largely shaped by school staff within some broad guidelines. Or, a head-to-head study might look at a reform that developed its own curricula versus one that worked with existing curricula. If done well (for example, there are several models of each type in the study), it might be possible to generalize findings to models of similar types. (For an example of such a study comparing the effects of alternative school improvement approaches, see Rowan, Correnti, Miller, and Camburn, 2009.)

[This page was intentionally left blank for double-sided printing.]

References

Alabama Department of Education, (2010). State of Alabama: Race to the top application. Montgomery, AL: Alabama Department of Education.Retrieved from https://www.alsde.edu/general/RACE-TO-THE-TOP.pdf

Aladjem, D. K., & Borman, K. M. (2006). *Examining comprehensive school reform*. Washington, DC: Urban Institute Press.

Choi, K. (2009). *Multisite multiple-cohort growth model with gap parameter (MMCGM): Latent variable regression 4-level hierarchical models*. IES unsolicited grant annual report.

Danielson, Charlotte (2007). Enhancing Professional Practice: A Framework for Teaching (2nd Edition). Association for Supervision and Curriculum Development.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 73*(3), 433–479.

Goldschmidt, P., Choi, K., Martinez, F., & Novak, J. (under review). Using growth models to monitor school performance: comparing the effect of the metric and the assessment. *Educational Measurement: Issues and Practice.*

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., and Darwin, M. 2008). *Turning Around Chronically Low-Performing Schools: A practice guide* (NCEE #2008-4020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37,* 351–360.

Patz, R. J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: Council of Chief State School Officers.

Pianta, Robert C., Karen M. LaParo, and Bridget K. Hamre (2008). Classroom Assessment Scoring System™ (CLASS™) Manual. Brooks Publishing.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and non-linear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Rowan, Brian, Richard Correnti, Robert J. Miller, and Eric M. Camburn (2009). *School Improvement by Design: Lessons from a Study of Comprehensive School Reform Programs.* Consortium for Policy Research in Education. Retrieved 5/26/2010 from http://www.cpre.org/images/stories/cpre_pdfs/sii%20final%20report_web%20file.pdf.