



Center for
Educator Compensation
Reform

Program Evaluation for the Design and Implementation of Performance-Based Compensation Systems

*Guide to Implementation:
Resources for Applied Practice*

Peter Witham

University of Wisconsin — Madison

Curtis Jones

University of Wisconsin — Madison

Anthony Milanowski

Westat

Christopher Thorn

University of Wisconsin — Madison

Steven Kimball

University of Wisconsin — Madison

Table of Contents

Introduction to Program Evaluation for the Design and Implementation of Performance-Based Compensation Systems	1
1: What Is the Program Being Evaluated?.....	3
Theories of Action	4
Developing a Logic Model From the Theory of Action	7
2: Developing Evaluation Questions	12
Evaluation Questions in Formative and Summative Evaluations	12
Developing Context Questions.....	12
Implementation Questions.....	13
Stakeholder Engagement and Communication	13
Program Fiscal Sustainability	16
Moving From Implementation to Outputs and Outcomes.....	17
Using the Findings From Evaluation Questions for Formative Feedback.....	19
Unintended Consequences	20
3: Using Qualitative, Quantitative, and Mixed-Methods Approaches	21
Using a Qualitative Methods Approach	21
Cost-Effectiveness Analysis	24
4: Evaluation Selection Framework	25
Determining Rigorous Evaluation Selection Frameworks.....	25
Designs for Answering Ultimate Outcome Questions.....	26
Non-experimental Designs	30
5: Disseminating Evaluation Results	32
Arranging Conditions to Foster Use of Findings	32

6. Managing TIF Program Evaluation Processes.....	35
Challenges of Managing TIF Evaluations	35
Choosing the Type of Evaluator.....	37
Strategies for Conducting a Successful Internal Evaluation	38
Strategies for Working with an External Evaluator	39
Using Meta-Evaluation in Both Internal and External Evaluations.....	42
Finding a Balance.....	42
Appendix 1 Internal and External Validity.....	43
Appendix 2 Joint Committee Standards	45
Appendix 3 Chicago Evaluation	
—Randomized Control Trial with Quasi-Experimental Matching	48
Appendix 4 Ohio Evaluation	
—Quasi-Experimental and Comparative Case Study.....	49
Appendix 5 Philadelphia Evaluation	
—Quasi-Experimental and Comparative Case Study.....	50
Appendix 6 Pittsburgh Evaluation	
—Quasi-Experimental Design with Implementation Analysis.....	51
Appendix 7 Power/Causality/Feasibility Analysis	52
Bibliography	53

Introduction to Program Evaluation for the Design and Implementation of Performance-Based Compensation Systems

The U.S. Department of Education (ED) requires all states, districts, and schools that receive funding through the Teacher Incentive Fund (TIF) to conduct an evaluation of their programs. This evaluation serves to ensure that grantees are: (1) following all Federal regulations and guidelines, (2) using formative implementation and outcome data for program improvement, and (3) using summative outcome data for accountability.¹ TIF program evaluations additionally contribute to the broader field of research on performance incentives and to policy discussions on this educational innovation. TIF evaluations are important for states and districts seeking to understand the relative benefits and costs of TIF programs. Given the current lack of national and international studies, particularly those using rigorous methodologies, on performance-based compensation systems, TIF evaluations provide important information to states and districts seeking to understand the relative costs and benefits of these systems.

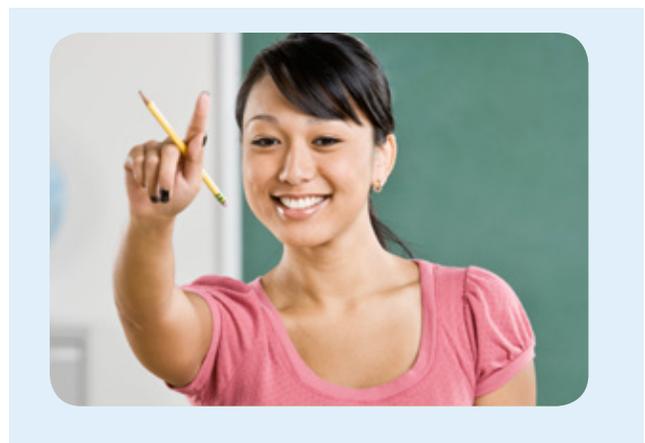
This guidebook provides strategies and resources to address the many complexities and challenges involved in evaluating a TIF program. The guidebook addresses the importance of rigor and professionalism in each of the stages of TIF evaluations, including conceptualizing, designing, conducting, and reporting. It begins by providing a process for identifying the logic of how the TIF program will

lead to the desired outcomes (Section 1), moves into how to develop evaluation questions that examine this logic (Section 2), and then explores methods for measuring these evaluation questions (Section 3) and choosing an appropriate evaluation selection framework (Section 4). The guidebook also provides best practices for disseminating evaluation findings (Section 5) and processes for choosing the right evaluator (Section 6).

Below is a brief overview of the guidebook.

Section 1: *What Is the Program Being Evaluated?*

This section provides both program designers and evaluators with resources for the conceptualization phase of the evaluation. Specifically, the section clearly articulates how TIF program designers must follow certain processes in order to meet their intended program goals. For example, the section



¹ TIF RFP: <http://www.ed.gov/legislation/FedRegister/announcements/2006-4/111406a.pdf>

provides guidance on how TIF program designers should identify a theory of action that helps shape a logic model. The logic model will then help TIF program designers visually depict how selected program components will lead to the desired outcomes. Clarifying the theory of action and depicting it in a logic model allows the TIF evaluator to construct appropriate evaluation questions to establish whether the program is accomplishing its goals. This section provides examples of common theories of action at work in TIF programs and strategies for making this theory concrete through a logic model that includes program inputs, activities, outputs, and short-, medium-, and long-term outcomes.

Section 2: *Developing Evaluation Questions.* Section 2 provides strategies for how program staff and evaluators can use inputs, activities, and outcomes represented in a logic model to create targeted, formative and summative evaluation questions. It addresses how the evaluation questions structured around inputs, activities, outputs, and short- or medium-term outcomes are most applicable to formative (periodic) evaluation and how questions about long-term outcomes are most applicable to summative, end-of-grant-cycle evaluations.

Section 3: *Using Qualitative, Quantitative, and Mixed-Method Approaches.* The third section addresses the appropriate application of qualitative, quantitative, and mixed-method approaches for measuring different aspects of the TIF program. It also examines how an evaluator can use specific evaluation questions to decide which of these approaches to use in which situations. This section encourages evaluators to use a balance of qualitative and quantitative approaches to examine each of the inputs, activities, context, outputs, and short- and medium-term outcomes within a TIF program.

Section 4: *Evaluation Selection Framework.* This section focuses on helping evaluators determine an appropriate selection framework for their evaluation; options include Experimental,

Quasi-Experimental, and Non-experimental frameworks. This section discusses each framework's requirements and strengths to establish causal relationships between an intervention and an outcome. The section also addresses the importance of an evaluator's framework selection by focusing on how to select a framework that allows for both a rigorous summative analysis of long-term outcomes (program impacts), and adequate information on outputs and short-term outcomes for formative use.

Section 5: *Disseminating Evaluation Results.* Section 5 provides best practices on how evaluators can disseminate their evaluation results to stakeholders. This section emphasizes that it is important for evaluators to communicate effectively with stakeholders throughout the evaluation because stakeholders must understand formative and summative evaluation results to make informed decisions about how best to improve programs. Furthermore, this section provides evaluators with helpful strategies for communicating evaluation results. These strategies include arranging conditions to foster use of findings, providing interim feedback, and providing standards for the preparation and delivery of formative and summative reports.

Section 6: *Making TIF Program Evaluation Processes* This final section guides TIF recipients through the process of developing evaluative systems that ensure objective, high-quality evaluations. It addresses the importance of choosing the right person to conduct the evaluation and outlines the necessary decisions that a project director should make in choosing who will conduct the formative and summative evaluations of the TIF grant. The section also discusses the potential ramifications of making uninformed decisions and explores how these decisions affect both the actual and perceived integrity of the evaluation. The section concludes with a discussion of how to promote appropriate relationships between internal and external evaluators and program staff, as well as strategies for developing Requests for Proposals, contracts, and budgets.

I | What Is the Program Being Evaluated?

The fundamental question that evaluation designers should ask is, “What is the program that is being evaluated?” Answering this basic question appears simple enough, but in reality, it involves a good deal of complexity and demands a systematic approach. With this in mind, this paper provides TIF program designers and evaluators with resources that they can use in the conceptualization phase of the evaluation. This section focuses on identifying the theory of action behind a TIF program and developing a logic model that depicts how the selected program components can lead to the desired outcomes. Thus, this section examines theories of action and the logic models used to put these theories into practice through a series of examples that TIF program designers and evaluators will find helpful as they work together to structure their own program designs and evaluations. Theories of action demonstrate how activities lead to outcomes, or in this instance, how educator incentives can lead to desired program outcomes such as higher student achievement.

TIF program designers often begin with a fairly clear idea of what the incentives and measures of performance will be in the program. Designers also have a general notion of what outcomes they hope to produce (such as higher student achievement), but lack an explicit understanding of how program activities lead to these desired outcomes. Thus, it is important for designers and evaluators to work together to develop an explicit theory of action. By developing a theory of action and creating a logic model from it, program designers and evaluators must think specifically about how program activities should lead to outputs and outcomes and provide a framework for further specifying evaluation

questions. A good logic model framework prompts evaluators to ask about each link in the causal chain from input to outcome. If the evaluators structure the plan properly, it will trace program effectiveness at each link, thereby providing formative information about the program’s impact and implementation fidelity.

A program’s theory of action is attempting to determine “What is the particular program supposed to do, and how is it supposed to do it?” The basic logic behind all performance-based compensation systems (PBCS) is that incentives have the potential to alter educator behavior, which will ultimately lead to increases in student achievement. More specifically, the logic is that educators make behavioral choices based on their perceptions and beliefs about how the incentives will affect them. Examples of this behavior may be increased effort in the classroom, gaining additional knowledge and skills from professional development offerings, or moving to



It is important for designers and evaluators to work together to develop an explicit theory of action.

schools where performance incentives are available. Articulating the particular logic behind the PBCS allows the evaluator to ask questions about whether what is “supposed to happen” is actually happening. For example, if the grantee bases the program on the logic that opportunities for bonus pay increase motivation and effort, then the evaluator would want to measure whether motivation and effort are actually increasing over the course of the evaluation. Consequently, to determine whether incentives affect behavior, an evaluation needs to examine educators’ perceptions and beliefs about the incentive system. Collecting this information is particularly useful when a program does not appear to be affecting educator behavior. The lack of a motivation effect is often due to a lack of understanding about how the incentives work or to perceptions that the incentive is too small to compensate for the effort required to meet performance goals.

This section specifically looks at three examples of theories of action: motivation theory, differential attraction and retention theory, and the theory of action for teacher evaluation. The section then examines logic models that build upon theories of action by showing the interconnectivity among program inputs, activities, outputs, and outcomes. To this end, this section presents two logic model templates and discusses their limitations and usefulness.

The process of developing a theory of action begins with the review and approval of the TIF grant application by the Department of Education (ED). The application describes specific elements of the program required by the Request for Proposal (RFP). While the substance and emphasis of each TIF program varies, the RFP requires each potential grantee to

1. Establish differentiated levels of compensation based on student achievement gains at the school or classroom level;

2. Conduct classroom evaluations multiple times within a single school year;
3. Ensure fiscal sustainability of the PBCS;
4. Develop an integrated approach to improving the educator workforce; and
5. Provide educators with incentives to assume additional responsibilities and leadership roles.

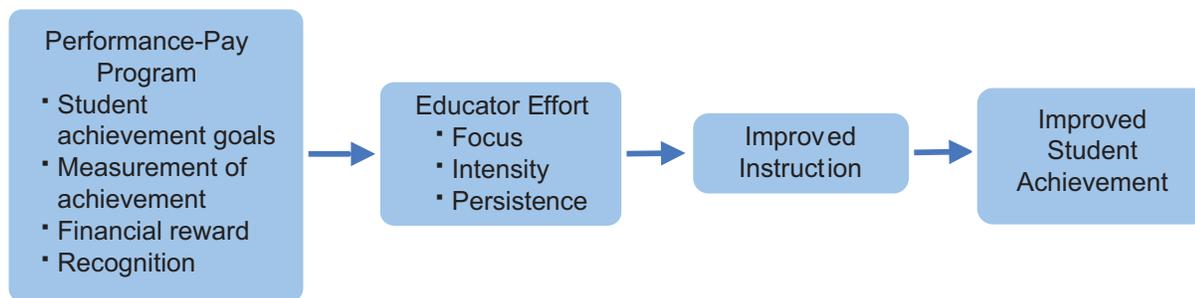
Evaluators use these RFP requirements as key pieces in designing the evaluation. In addition, evaluators collaborate with TIF staff to review other important documentation, such as communication plans, implementation plans, and progress reports to inform the evaluation design.

Evaluators may find it useful to interview TIF program designers and administrators, asking them to describe program elements in detail, give their opinions on which aspects of the program should receive the most attention, and specify the outcomes they expect to see. Importantly, evaluators should be sure to ask TIF program staff how they expect these program elements to influence participating educators to make changes that will lead to desired outcomes. Evaluators will find this kind of question helpful because it identifies the intermediate actions or results that would have to take place to cause the intended outcomes. For instance, if a TIF program’s desired outcome is improved student achievement, it is likely that teachers will need to change some aspects of their instruction or that schools will need to change how they support instruction, perhaps by allocating more time for instruction in tested subjects or acquiring different curricular materials.

Theories of Action

TIF program staff and evaluators may find it useful to develop a diagram that represents the theory of action and its associated causal chain. Described earlier, theories of action provide program staff and evaluators with specific ways in which TIF

Figure 1.1: Motivation theory of action



activities lead to programmatic outcomes. A causal chain illustrates the expected connection between programmatic features (e.g., incentives for growth in student achievement) and desired outcomes. Most TIF programs can modify these generic examples given below to fit their requirements.

One generic theory of action that could fit TIF projects is motivation theory (Figure 1.1). The premise of motivation theory is that incentives motivate educators to modify their behavior in ways that make them more likely to receive the incentive, and that these modifications will lead to improved student achievement.

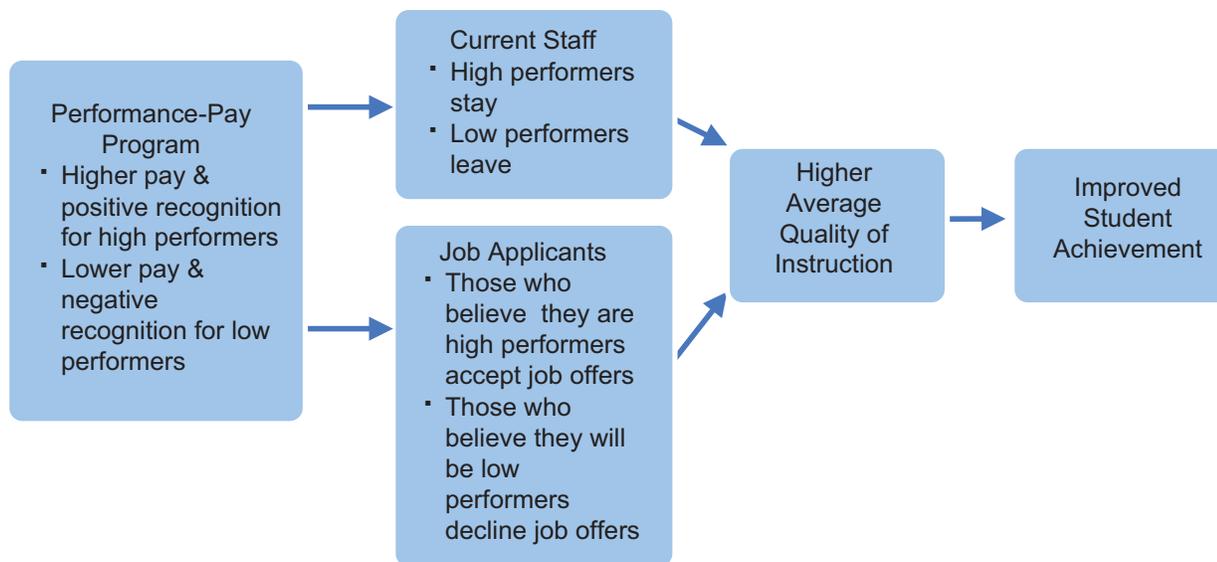
The specific program elements, or “active ingredients,” of the program include performance goals that identify the levels of student achievement, during a specified period of time, that are required for educators to receive financial rewards. In addition to the financial award, educators receive school or district recognition as a reward for their exemplary performance. Because of these rewards’ desirability, in theory, educators will want to pursue the student achievement goals. Educators would do this by changing the focus of their efforts, changing their effort level, and sustaining this enhanced effort to achieve the goals. For example, teachers might be motivated to focus their instructional time on covering tested content, to find and use instructional practices that are likely to improve

student learning, or to persist in working with struggling students. These efforts, in turn, would then contribute to improved student achievement. This theory of action is a particularly good fit for programs in which a set of prespecified bonus payments is associated with various levels of improvement in student achievement, be they at the classroom, team, grade, or school level.

Another generic theory of action, differential attraction and retention theory, provides a good fit for programs that reward a prescribed quota of educators based on a relative measure of student achievement (Figure 1.2). For example, the program provides a bonus to teachers whose classrooms are in the top 20 percent on a value-added measure of student achievement.

In differential attraction and retention, the active ingredients of the program are the bonuses that provide higher pay for educators who are high performers. Some form of public recognition, showing that winning teachers are performance exemplars, typically accompanies the bonus. Teachers who do not earn the bonus receive lower relative pay and no recognition as high performers. In programs that make public the names of teachers with below-average performance metrics, these educators are marked with the stigma attached to low performance. These consequences have effects on both current staff and on teachers who apply for

Figure 1.2: Differential attraction and retention theory of action



Source: Adapted from Milanowski, 2002.

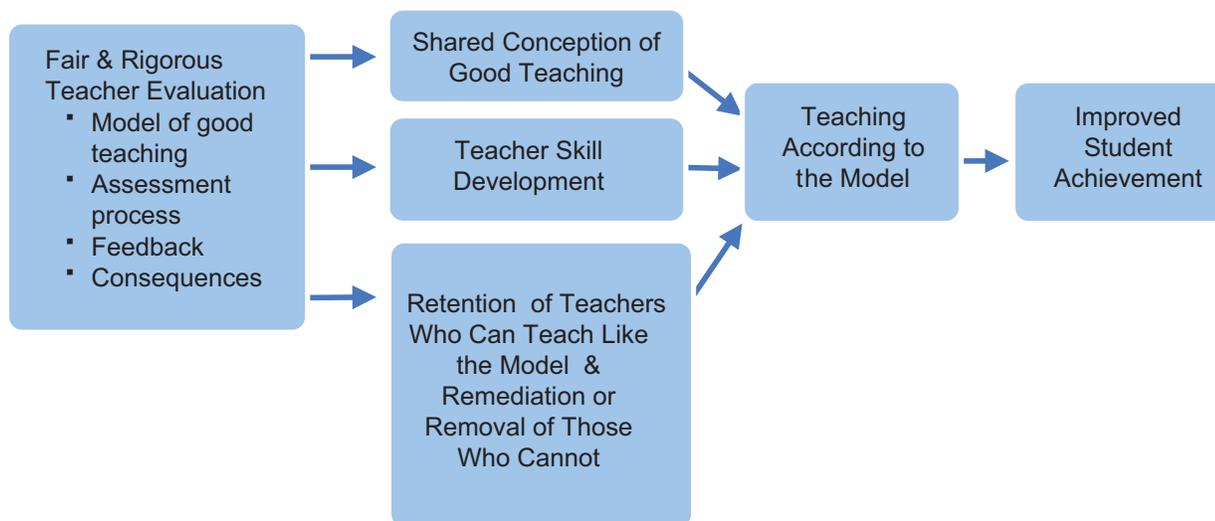
open positions—high performers are more likely to stay with the district, and poor performers are less likely to stay. This effect may be small in any given year, but over time, the differential attrition could improve the average quality of instruction. The larger the performance bonus, and the smaller the “automatic” base pay increase provided via the traditional pay schedule, the stronger this effect will be.

Bonuses and recognition also can affect who initially applies for positions in the school or district. If job candidates are aware of the bonus program, their willingness to apply will likely depend (in some part) on an internal assessment of whether they have the professional skills and the philosophical commitment to succeed. Those who do not have these skills or commitment are more likely to self-select out of the hiring process, and those that do have them are more likely to apply. Over time, the school or district is likely to gather a teaching force that tightly aligns with its theory of action, thereby improving instructional quality and raising student achievement.

A third theory of action, theory of action for teacher evaluation, involves the effects on student achievement of a fair, rigorous, and objective evaluation process (Figure 1.3).

This theory requires that an evaluation process designed to improve teaching begin with (1) a set of performance rubrics that outline a model of good teaching, (2) a fair and rigorous assessment process that results in an accurate portrait of a teacher’s performance, (3) feedback to teachers about how they performed, and (4) consequences for good or poor performance. When teachers implement these processes effectively, they develop a common understanding of good teaching that sets the stage for a culture of high expectations for instructional quality and develops their instructional skills accordingly. In this theory of action, as teachers develop a shared conception of good instruction and hone their skills to teach in ways consistent with the rubrics, students experience teaching that is more effective. So long as the model does, in fact, capture those elements of instruction that drive student learning, students should improve academically. To obtain

Figure 1.3: Theory of action for teacher evaluation



these high outcomes most quickly, districts should recognize and/or pay more for high-performing teachers and should identify low-performing teachers for remediation or termination.

One or more of the theories of action described above will fit many of the existing TIF projects, and grantees can combine or modify them to meet individual needs. Importantly, all of the theories postulate that improved classroom instruction increases student achievement. These theories are consistent with current research that shows that teachers and classroom instruction are a strong school-level factor that determines student achievement (Goldhaber, 2009; Rivkin, Hanushek, and Kain, 2002). With these theories in mind, evaluators may want to measure the effectiveness of a TIF project based on whether instruction or instruction-related behavior has changed in response to incentives.

Developing a Logic Model From the Theory of Action

Logic models are helpful tools that build on the theory of action. The name “logic model” emphasizes that the goal is to depict the program’s causal flow (i.e., how committing a set of inputs should lead to a set of desired outcomes, through specific

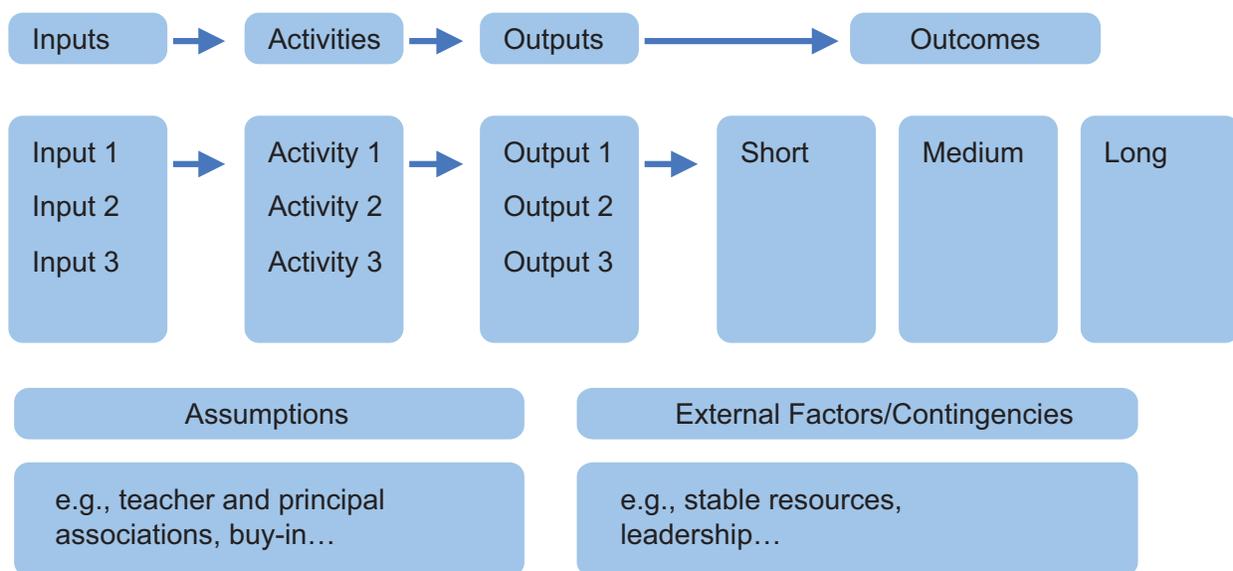
activities and their associated outputs). The value of a logic model is its clear representation of the theory of action, and the connections among program inputs, activities, outputs, and outcomes (described below). By visually depicting the causal chain, an evaluator can begin to think about how to construct evaluation questions that will answer whether the program is “doing what it is supposed to do.” As discussed below, a logic model can also capture contextual factors outside of the program, such as political or fiscal issues, that may have a strong influence on the level at which the inputs and activities affect the outputs and outcomes. One common logic model template, presented in Figure 1.4, lists inputs, activities, outputs, and outcomes in columns from left to right. The figure does not show any causal interconnections among the short-, medium-, and long-term outcomes. However, these important interconnections are discussed later in the paper.

Working from left to right in Figure 1.4:

- Program inputs are the resources the program uses to start and sustain it. TIF project inputs might include the funds provided by the federal government, the

Figure 1.4: Logic model framework

Logic Model Shell



Source: Derived from Fretchling, 2007.

- project staff hired with these funds, and the support of important stakeholders.
- Program activities are the tasks and operations that program staff and others engage in to achieve program goals. Central to most TIF programs are activities such as designing the incentive structure and measurement systems, creating a communication plan and communicating that plan to educators, measuring student achievement, and making payouts.
 - Program outputs are the direct results of program activities, services, and products. Fretchling (2007) theorizes that outputs are the most immediate indicators that the theory of action is working, and advises evaluators to identify at least one output for each activity. In most TIF projects, outputs will include communication products (e.g., websites, program brochures), reports

on school- or classroom-level student achievement or growth, reports on teacher and principal observations, and, most important, the correct educators receiving the proper incentive payout.

- Program outcomes are the results of these activities, such as increased student achievement. Often, logic models distinguish between intermediate outcomes, such as changes in beliefs or behaviors related to instruction, and ultimate outcomes, such as improved student achievement.

Logic models may also include important contextual factors that can influence how strongly the inputs and activities affect the outputs and outcomes. At most TIF sites, important contextual factors may include other programs or initiatives aimed at improving instruction or achievement (e.g., new professional development programs, new curricula), resource sufficiency or shortfalls, and

school or school day organization. These alternative elements can have a wide array of effects on TIF programs. For example, external initiatives can send competing messages to teachers about pedagogical priorities; budget cuts may prevent the hiring of new teachers; and the school day may be too full to allow evaluators to provide teachers with post-observation feedback. Contextual factors can therefore limit the impact incentives can have or can augment them, making it difficult to attribute a change in behavior or achievement to the incentive program. Thus, evaluators should identify potential contextual influences and include them in the logic model. For example, the logic model framework shown in Figure 1.4 asks evaluators to examine each link in the causal chain from input to outcome. Thus, if the evaluators structure the evaluation plan properly, it should trace the program's effectiveness at each link, thereby showing the program's impact.

As is further discussed below, such an evaluation will also provide evidence that directly addresses causal claims regarding the TIF program and desired long-term outcomes. This is important, as there can be multiple causes for a desired outcome, including contextual factors and other programs that the local or state education agency (LEA or SEA) is implementing. The logic is that if the evaluator observes the intended long-term outcomes, he/she will be more confident attributing them to the program, provided that the evaluation shows that program activities were performed as intended, produced the outputs intended, and that these outputs produced the short-term outcomes expected. The evaluator would also have more confidence in the evaluation results if the program designer takes these factors into account in the logic model. These contextual factors could include changes in program leadership, changes in student populations, and concurrent changes in instructional programs.

As stated in the beginning of this article, translating a theory of action into a logic model requires evaluators and program designers to work together.

This work may require evaluators to interview administrators, read program documentation, and meet with program designers and administrators to establish how outputs affect outcomes. Evaluators also need to be flexible about the components of the logic model. For example, as the TIF program evolves, changes may occur.

One example of a completed logic model, displayed in Figure 1.5, incorporates the combined effects of a schoolwide incentive and a more rigorous teacher evaluation system. The model assumes that incentives are based on attaining schoolwide student achievement goals (either attainment or improvement), and as will be explained, the incentive and evaluation components have the potential to reinforce one another and exponentially affect instruction.

This model includes multiple inputs, such as (1) federal TIF and local funds used to finance the project, (2) project staff, (3) district leadership, and (4) key stakeholder (e.g., school board, teacher union) support. Once these resources are in place, administrators can carry out a set of activities that begin with program design and continue with the disclosure of important program features. Assuming a school-level performance incentive, administrator activities would also include setting school performance goals, defining how school performance will be measured, outlining the payouts for various levels of performance, and determining whether high performance will be recognized in other ways. With respect to the particular component of teacher evaluation, for example, communication could include standards of teacher performance (a model of "good" teaching), teaching performance measures (e.g., number and timing of observations, who will observe), and how the school will recognize good teacher performance and remediate poor teacher performance. Note that if educators do not understand the activities within a logic model, then the incentive will not likely motivate them.

Additional key TIF program activities involve measuring performance, providing feedback, and delivering consequences to teachers and/or principals. When considering the performance incentive component, administrators must assess student achievement, calculate changes in attainment or value added, calculate payout amounts, and prepare and distribute incentive checks. These activities are the most basic mechanisms of an incentive program, and most evaluations need to determine whether they occurred as intended and how well they were completed. The teacher evaluation component requires that administrators observe classrooms, provide feedback to teachers about how well they are performing, and recognize teachers who are performing well and identify for remediation those who are performing poorly.

Most evaluators will also be concerned with whether TIF program staff implemented the activities and realized the intended outputs. This is the basis for assessing implementation fidelity; whether TIF program staff members enact an intervention as they intended. From a formative perspective, most program administrators will want to know as soon as possible if the intended activities are taking place and producing outputs so that administrators can take corrective action if they are not. From a summative perspective, an assessment of implementation fidelity is important because gauging accuracy must precede any conclusions about program design, whether they are positive or negative.

For example, Figure 1.5 shows TIF program outcomes, including the short-, medium-, and long-term effects, that can occur from the hypothetical TIF project. The first short-term outcome listed is that teachers are motivated to change their behaviors around instruction in an attempt to receive the incentive. This change in behavior, a short-term outcome, is important because one of the fundamental ideas behind performance incentives is that they will motivate educators to change their behaviors in ways that improve student learning.

This logic model also postulates that teachers will be motivated to change practice by the teacher evaluation component.

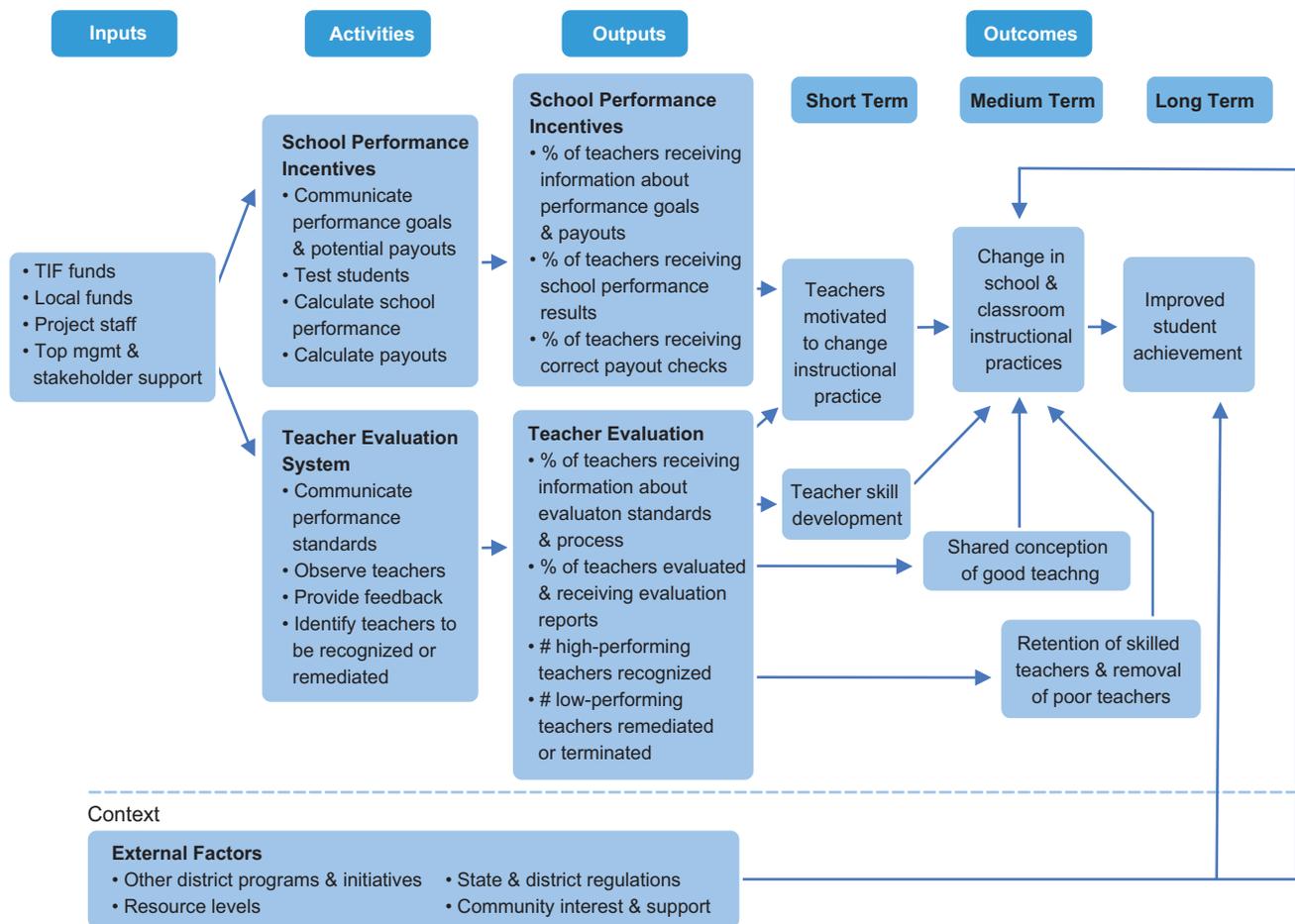
In Figure 1.5, the diagram also shows that the first medium-term outcome is that teachers make changes—most importantly, to instructional practice. This logic model assumes that teachers primarily affect student achievement via their instruction. Therefore, evaluators should examine instructional change when conducting a comprehensive evaluation.

In addition, Figure 1.5 illustrates that another medium-term outcome is that the evaluation process will lead to a shared contribution of good teaching. More specifically, this outcome may represent a culture change in some schools, if the performance standards are more rigorous, and prior evaluation practices were lax. By defining what it means to be a good teacher, this shared conception in turn reinforces changes in practice that are consistent with the performance standards through peer pressure and teachers' desire to fit in with the school culture.

Further, Figure 1.5 demonstrates that an additional medium-range consequence is that the teacher evaluation component leads to the retention of skilled teachers and the removal of poor performers. This outcome raises the average level of instructional practice. In addition, this outcome overlaps the medium- and long-term categories because these effects should continue over the life of the TIF program.

Figure 1.5 also illustrates that the long-term outcome is improved student achievement. Most TIF programs expect improved student achievement to be the “bottom line” impact. While most evaluators would assume that the major influence on improved student achievement is changes in instructional practice, the logic model also recognizes that contextual factors (such as programs

Figure 1.5: Logic model for incentive program involving school incentives and teacher evaluation



outside of the TIF activities) are likely to have an important influence.

Figure 1.5 includes contextual factors to remind us that an evaluation needs to identify their influences when making judgments about the impacts of TIF program components. For example, a new curriculum could reinforce the effect of the incentive program, if it aligned more to state test content and to the instructional expectations underlying classroom observations. On the other hand, a major curriculum change could also work against the incentive if it competed for teachers’ attention or took time away from activities that would be more productive for student achievement. Including important contextual features in the logic model reminds evaluators to be on the lookout for complicating or countervailing effects.

In conclusion, this section has shown the importance for evaluators to use the theory of action to develop a logic model that will guide evaluation questions that focus on causal links between major program elements and ultimate outcome goals, such as improved student achievement. Further, Section 1 has illustrated the value evaluators gain by constructing a logic model that specifies the TIF program inputs, activities, outputs, and short-, medium-, and long-term outcomes. The logic model will help evaluators to illustrate the key features, or “active ingredients,” of the TIF program and connect them to outcomes throughout the life of the TIF grant and beyond.

2 | Developing Evaluation Questions

This section builds on Section 1 of the *Evaluation Guidebook*, which examined the theory of action behind a TIF grant and its representation in a logic model. This section provides strategies for program staff and evaluators to develop formative and summative evaluation questions using the inputs, activities, outputs, and outcomes represented in a logic model.

Evaluation Questions in Formative and Summative Evaluations

Typically, evaluations of programs such as TIF include both formative (periodic) and summative (end of the grant cycle) evaluations. Formative evaluations focus on answering implementation-oriented questions (inputs, activities, outputs, and short-/medium-range outcomes) using both qualitative and quantitative methods. Program staff can use these evaluation results to provide ongoing feedback about program implementation for potential areas of improvement. On the other hand, summative evaluations tend to connect to long-term outcomes. While summative evaluations rely upon formative information to address issues of treatment fidelity, their primary purpose is to judge the overall effectiveness of a program. Overall, both the questions answered in formative and summative evaluations are important, and if evaluations are to be comprehensive, they should include both.

Program designers and/or administrators and the evaluators must develop evaluation questions for formative and summative evaluations collaboratively. The most useful evaluations incorporate questions that the program staff finds important. It is worth taking the time at the beginning of an evaluation to have both groups review the logic

model and brainstorm questions. Evaluators can then propose a final set of questions based on resources available, Department of Education requirements, and expected program implementation issues. Agreement on a working logic model and on evaluation questions derived from it also helps both groups feel comfortable with the evaluation work. Program administrators are more confident that the evaluator understands the program, and evaluators are more confident that they know what is expected. Agreement between evaluators and program staff on final questions sets the stage for a successful project.

Developing Context Questions

One important area for an evaluator to consider (particularly for formative evaluation) is the contextual factors that could affect implementation. Evaluators can identify some of these influences up front and address them with evaluation questions, but evaluators will likely only hear about some of them from educators during the evaluation. Thus, the evaluation design may need to incorporate at least some evaluator interviews with educators to ask general questions about other influences they feel might affect instruction and efforts to improve student achievement and whether these influences reinforce or distract from the incentive program. Examples of contextual questions that evaluators can build into the evaluation design follow:

- Has the district (or other governing body) initiated any other programs that could affect teaching or leadership in schools?
- When did the district initiate these programs?

- What did they require of the educators covered by the incentive program?
- Have state policies or procedures involving student testing changed (e.g., content and timing of tests, which grades are tested)?
- Are there state or district regulations or policies that affect classroom observations? (For example, is a particular evaluation process required, and how does it relate to the process used in the TIF program?)
- Are there state or district regulations or policies that affect the incentive payout? (For example, are any bonuses included in retirement benefit calculations? How much is the take-home amount from a bonus reduced by tax withholding?)
- Have other district programs supported or conflicted with program activities?
- Has the general level of funding affecting TIF program schools increased or decreased?
- If funding levels have changed, has this prompted additional hiring or layoffs of educators?

Implementation Questions

Some of the most important questions an evaluator must consider (*for both formative and summative evaluations*) are whether projects have implemented inputs and their associated activities effectively. As discussed in Section 1 on theories of action and logic models, each performance incentive system includes numerous inputs. Some of the most important inputs include:

- Plans for stakeholder engagement and communication;
- Award structure;
- Plans for fiscal sustainability.

Each of these inputs has activities associated with it that the grantee must implement in order for the program to have the highest likelihood of producing

the desired output and, ultimately, outcome. A formative evaluation will align these programmatic activities with evaluation questions to determine whether grantees are implementing the inputs most effectively.

The following section describes the process of aligning inputs and associated activities with evaluation questions and provides a framework for conducting a formative process evaluation. This section also addresses how an evaluator moves from evaluation questions to measurement and outlines a number of instruments that determine whether districts are implementing a program effectively. Additional information about these inputs and associated activities is on the Center for Educator Compensation Reform (CECR) website.²

Stakeholder Engagement and Communication

Two key inputs in a performance incentive system are stakeholder engagement and a communication plan. The appropriate engagement and systematic communication with stakeholders is crucial to the successful design and implementation of a TIF grant. Key stakeholder groups must participate in developing the system and accept it if it is to be sustainable in the larger community and individual schools. Districts' implementation of performance-pay systems shows that securing stakeholder (particularly teacher) buy-in and commitment to a new compensation system is essential to a program's success and its long-term sustainability. Tables 2.1 and 2.2 below show stakeholder engagement and communication inputs, activities, and evaluation questions.

Example—Stakeholder Engagement Activity: One important activity in engaging stakeholders is constructing a compensation committee. A working committee that is representative of major stakeholders gives the plan a better chance of succeeding by engaging school and district leaders, including

²<http://www.cecr.ed.gov/pdfs/guide/CECRchecklist.pdf>

teacher union and association representatives, from the outset. The evaluator should measure the implementation of this activity, linking it to a number of evaluation questions. For example, the formative evaluation could ask whether the compensation committee includes a diverse set of educators, policymakers, and practitioners that the performance incentive system directly affects.

Table 2.1 illustrates two activities in a stakeholder engagement plan and associated evaluation questions that frame the measurement of the implementation of this plan. Document analysis, interviews, surveys, and focus groups are some of the instruments that measure the alignment of these activities with the standards represented in the evaluation questions. For more information on developing a representative committee and other important activities when developing a stakeholder engagement plan in a TIF grant, see *Stakeholder Engagement and Communication Guidebook Module*.³

Example—Communication Activity: To ensure smooth implementation, designers of performance incentive programs must also keep all stakeholders informed about the program and its components. Program designers must write clear communication plans that include a range of activities. Aligning evaluation questions with these activities will allow the evaluation team to determine whether the SEA or LEA is implementing the communication plan, and whether it is having its desired impact on stakeholders.

For example, the formative evaluation could measure whether the representative committee and district leadership responsible for program implementation clearly articulated how they intend to communicate details of the new compensation plan to various stakeholders, the methods they will use, who will be responsible for developing and communicating information, and timelines for implementation. Table 2.2 illustrates a number of key activities within a communication plan and associated evaluation questions for measuring plan implementation. From this point, an evaluator can use a number of instruments to measure the alignment of these activities with the standards represented in the evaluation question. For more information on developing a representative committee and other important activities in a communication plan for a TIF grant, see *Stakeholder Engagement and Communication Guidebook Module*.⁴

Award Structure

The award structure articulates the criteria for receiving an award. The activities associated with this input include determining who is eligible for the incentives, the criteria grantees will use to determine whether they receive an award, and the size of the award. Table 2.3 provides more information on award structure inputs, activities, and evaluation questions.

Example—Award Structure Activity: One of the most important award structure activities is determining

Table 2.1: Input—stakeholder engagement

Activity	Evaluation question
Construct representative stakeholder group	Have the districts assembled a compensation committee that includes school district officials as well as teachers and others (e.g., principals) whose salaries the new plan will affect?
Engage representative stakeholder group	Have the districts invited individuals and groups to serve on the compensation reform committee so that they are active participants in discussions, planning, and decisions from the beginning (superintendent, teachers union, representative group of teachers, principals)?

³<http://www.cecr.ed.gov/pdfs/guide/CECRStakeholderEngagement.pdf> <http://www.cecr.ed.gov/pdfs/guide/CECRStakeholderEngagement.pdf>

Table 2.2: Input—communication plan

Activity	Evaluation question
Establish clear award criteria	Have the districts developed a communication plan that clearly explains to teachers, principals, and others possibly affected by the performance-pay plan what the criteria to determine eligibility for a performance award are and what they must do to earn one?
Establish clear connection to additional programs	Have the districts developed materials that clearly explain professional development opportunities for teachers and principals desiring to improve their performance so that they can earn a performance award?
Develop strategy for informing district leadership	Have the districts articulated steps for informing district-level leadership across the range of departments likely to be involved in some aspect of the compensation plan about the details of the plan components?
Develop multiple means for distribution of communication	Have the districts determined multiple means for distributing information to educators and the public (e.g., brochures, pamphlets, newsletters, town meetings, e-mail alerts, and an updated website)?
Develop multiple ways to access information	Have the districts determined alternative means by which educators can gather information quickly and easily (e.g., confidential hotline, convenient after-school drop-in sessions, and trained individuals at each school site who can answer questions)?
Develop strategy for addressing media	Have the districts specifically and forcefully addressed the media (e.g., do they have a plan in place to respond to Freedom of Information requests—both with internal and external constituents)?
Develop strategy for targeted communication dissemination	Have the districts established targeted activities related to key events in the life cycle of the plan (e.g., program kickoff, specific measures of performance, the payout)?

Table 2.3: Input—award structure

Activity	Evaluation question
Determine eligible personnel	Have the districts decided which and how many educator positions will be included (e.g., all classroom teachers, only teachers of core academic subjects, paraprofessionals as well as teachers, assistant principals as well as principals)?
	Have the districts decided if individuals, groups, or both will receive awards?
	Have the districts determined which groups (e.g., all teachers in the school, all math teachers in the school, all fourth-grade math teachers in the school) will receive awards?
	Have the districts decided whether the new compensation plan will be voluntary or mandatory?
Determine measurement criteria for personnel	Have the districts determined how to appraise the performance of those who teach non-tested subjects and grades (e.g., preschool, art, music, physical education, fifth-grade science)?
	Have the districts determined whether the award structure will link directly to desired teacher behaviors and student outcomes?
	Have the districts determined whether to use competitive elements to determine how performance targets are established (e.g., average growth of student achievement in math is in top quartile of participating teachers)?
Determine what is included in the award	Have the districts determined the amount of the financial award?
	Have the districts determined whether to use any noncash awards (e.g., housing incentives, tuition assistance, tax incentives, and additional credit toward retirement)?
Determine timing of the award	Have the districts determined when to distribute the award?
	Have the districts determined whether to phase in the compensation plan as new teachers are hired, or will they transfer all teachers to the new plan at the same time?

how to measure the performance of the personnel included in the incentive system. If the evaluators are to measure the implementation of this activity, they must link the activity to a number of evaluation questions. For example, the formative evaluation could measure whether a process is in place for selecting the most rigorous and accurate measures of educator performance. Once the activity and evaluation question connect, the evaluator can use a number of instruments (case study, interview, survey, etc.) to measure the degree of input implementation. Table 2.3 illustrates a number of the activities and associated formative evaluation questions that measure the implementation of this input.

Program Fiscal Sustainability

Another important input for a performance incentive system is a plan for fiscal and programmatic sustainability. Teachers will not accept incentive programs nor will the programs be effective if teachers do not believe that state and district officials will actually deliver earned financial rewards as promised. No matter how well-designed a compensation system may be, and no matter how much organizational or political support it

has, it will not succeed if it is not affordable. Table 2.4 lists program sustainability activities and associated evaluations.

Example—Fiscal Sustainability Activity: One example of an important program sustainability activity is developing a plan for projecting costs of the performance incentive program. States and districts should project maximum program costs each year to avoid the possibility that the number of teachers or schools that qualify for awards exceed available funds. In order for the evaluator to measure the implementation of this activity, he or she must link the activity to a number of evaluation questions. For example, the formative evaluation could measure whether the state or district has collected and analyzed the necessary student achievement and teacher data to estimate probable financial exposure. Evaluators could use a number of instruments to measure the implementation of this input, including document analysis, interviews, surveys, and focus groups. For more information on cost projection and other important activities involved in program fiscal sustainability with TIF grants, see *Paying for and Sustaining a TIF Grant Guidebook Module*.⁵

⁵<http://www.cccr.ed.gov/pdfs/guide/payingFor.pdf>

Table 2.4: Input—fiscal sustainability

Activity	Evaluation question
Establish cost of award system	Have the districts calculated the maximum cost of the new compensation plan year by year?
	Have the districts identified revenue sources to pay for the new compensation plan?
	Have the districts purposefully constructed an overall plan to ensure long-term financial sustainability?
	Have the districts determined if their financial resources are adequate for assessing and maintaining data quality standards (e.g., uncovering data quality errors at the school level, administering data quality checklist at the school level)? Do the districts have a plan to account for the additional resources needed to implement and maintain data quality?
Establish timing of awards	Have the districts decided the frequency and timing of awards (e.g., one-time bonus; permanent increase to base salary; premium for teachers of hard-to-fill subjects in addition to their regular salary; in-kind payment made in the form of goods and services, rather than cash)?
	Have the districts decided how close to the period of performance awards will be paid?
Determine who will distribute awards	Have the districts specified an agency that will actually pay the awards (e.g., SEA, school district central office, or community foundation)?

The findings from the measurement of these evaluation inputs of stakeholder engagement and communication plans, award structures, and plans for fiscal sustainability and their associated activities will provide feedback for guiding a program through its startup phase, ensuring its ongoing quality, and improving it as it matures.

Moving From Implementation to Outputs and Outcomes

In addition to focusing on the implementation of inputs and their associated activities, evaluators must also ask questions about outputs and their relationship to short- and medium-range outcomes. Here, the emphasis should be on how the activities led to certain outputs, whether there is progress toward the short- and medium-term outcomes, and if not, what has gone wrong. Each of the inputs in a TIF program will lead to a number of outputs, which have an impact on the short-, medium-, and long-term outcomes. Using the input of a communication plan (see Table 2.2) as an example, evaluators could measure a number of outputs in relationship to the plan's implementation. These outputs could include:

- What percentage of educators received information about performance goals and payouts?
- Did they receive the information early enough in the school year to influence behavior?
- What percentage of educators received the results of school performance analysis?
- Where the reported results accurate and timely?

Short-Term Outcomes

The basic logic underlying TIF programs is that outputs like incentive payouts and classroom observations (and their attached consequences) will

motivate educators to change their practices. The logic continues that this motivation is dependent on educators' beliefs about the program, their abilities and resources to change, and their professional needs and values. Thus, many of the short-term outcome questions should be about how the program outputs affect educator beliefs and how these beliefs relate to motivation.

As a first step on this path, evaluators must determine if educators understand what the payouts and consequences are, what level of performance the payouts or consequences are contingent upon, and how performance is measured. Evaluators must also determine the degree to which educators accept that the performance goals and/or practice standards are appropriate for (i.e., legitimate, attainable, and consistent with) other goals set by the state, district, or school. Without such acceptance, teachers' motivation to work toward TIF-related goals is not going to be strong. Since motivation is not likely without such understanding and acceptance, it is especially important, in a formative evaluation, for evaluators to ask educators if these perceptions and cognitions have been the outcome of program outputs. Some potential evaluation questions about these short-term outcomes include:

- Do educators understand the performance goals, performance measures, and potential incentive payouts?
- Do educators understand the practice model or performance standards underlying the observation process, how the process works, and any consequences of the observations?
- Do educators accept the performance goals as worthwhile and attainable?
- Do educators accept the practice model or performance standards as an appropriate and attainable standard of practice?
- Do educators report that the potential incentive payout motivates them to change their behavior?
- Do educators report that the observations motivate them to change behavior?

- Do educators report that they pursue professional development related directly to their performance as evaluated by the teacher evaluation process?
- What behavioral (practice) changes do educators report being motivated to make?

However, evaluators should not limit evaluation questions regarding short-term outcomes to questions about attitudes or perceptions. They can also ask whether professional development records show an increased demand for courses related to practice change, whether teachers have requested professional development help, or whether educators ask for changes in schedules or more teaching resources.

Medium-Term Outcomes

In the majority of TIF programs, the key medium-term outcome is improved instructional practice, which should lead to improved student outcomes. The basic medium-term outcome evaluation question will thus center on whether instructional practices have changed. (Note that for school leaders, practice changes intended to support improved instruction would be the medium-term outcomes.)

In a TIF evaluation, the assessment of change in educator instructional practice can be extremely complex. An evaluator must take into account the degree of alignment between a district's vision of instruction and classroom observational standards or measures of student achievement. While an in-depth discussion of the important components of changes in educator instructional practice is beyond the scope of this guide, evaluators may want to consider adapting these generic questions about instructional practice:

- Are teachers spending more time on core or tested subjects?
- Have schools changed schedules to allow for more instruction time in core or tested subjects?

- Have teachers and schools aligned the curriculum to the tests or to underlying state standards?
- Have teachers and schools acquired and used new textbooks or other instructional materials related to tested subjects?
- Have teachers and schools increased their use of data to track student performance and identify struggling students?
- Have teachers and schools implemented specific interventions or extra assistance strategies to help struggling students?

Note that the important medium-term outcome evaluation questions will focus on those aspects of instruction that teachers and schools can control and on changes they can make in response to incentives. Thus, it is important for evaluators to distinguish between changes initiated by the state or district and those initiated by teachers and schools. Since the evaluator cannot completely anticipate these changes when designing questions, he/she should also include a general question about any other changes teachers or schools make in order to qualify for the incentives.

For programs that include a strong teacher evaluation component, especially those that base some part of the incentive on observations, the important evaluation question is whether teachers' practice is becoming more like the model used as the basis for the evaluations. In this case, the observation rubric or rating scale may provide a good place to start in developing specific questions about instructional practice change. In addition, evaluators may want to ask teachers and school administrators if they perceive that a shared conception of good teaching (based on the practice model underlying the evaluation system) has been developing.



The design of the evaluation is crucial in showing whether any change in student outcomes is attributable to the TIF incentive program.

Last, evaluators could ask teachers and administrators if the evaluation process was successful in remediating or removing poor teachers, whether teachers who received positive evaluations were more likely to stay, and whether those receiving poor evaluations were more likely to leave. Evaluators could also use district human resources information system data to ask whether those evaluated as better performers were more likely to stay, controlling for age, experience, and other factors known to influence teacher turnover.

Using the Findings From Evaluation Questions for Formative Feedback

The evaluation questions above that address inputs, activities, outputs, and short-/medium-term outcomes are most applicable for formative (periodic) evaluation. When evaluators use both qualitative and quantitative methods to answer each of these sets of questions, they are able to provide the program staff with a rich source of information for programmatic feedback. When evaluators deliver this information correctly (See Section 5, Disseminating Evaluation Results), it can have a significant impact on the program, as program staff can make adjustments to the inputs and associated activities to improve outputs and outcomes.

Long-Term Outcomes

Evaluators use long-term outcome questions primarily for summative evaluations, as they address the “bottom line” question of whether student achievement measures have, in fact, changed. There are several ways to think about these long-term changes in achievement, especially at the school level, including changes in average attainment, value-added productivity estimates, and improvements in value added. The latter is particularly interesting because the goal for changes in instruction is to improve classroom or school productivity. For high schools, where testing is more limited, it may also be useful for evaluators to think in terms of outcomes such as graduation rates, college or career readiness, and postsecondary participation.

As is discussed later in this guide (Section 4, Evaluation Selection Framework), the design of the evaluation is crucial in showing whether any change in student outcomes is attributable to the TIF incentive program. However, even if TIF project evaluators cannot use a strong design, they should at least track trends in outcomes and compare the trend after TIF implementation to the trend before implementation. District stakeholders will typically

want to know, at a minimum, if some change has taken place. Evaluators could make some judgment of impact even without a strong impact evaluation design if the evaluation has addressed all the elements of the logic model.

For example, if the evaluation has found (a) faithful implementation, (b) few contextual influences, (c) educators motivated by the incentive, and (d) changed instructional practice, then most decision-makers would be comfortable attributing some of any positive upward trend in student achievement to the incentive program. On the other hand, if the evaluation has found faithful implementation but no changes in achievement, this would be actionable evidence that the program had minimal impact.

Unintended Consequences

In addition to input, activity, output, and outcome questions, it is important for evaluators to develop specific questions about unintended or unexpected consequences of TIF programs. Research on performance incentives in the private sector suggests that there are common types of unintended consequences that can assist in the development of these evaluation questions for TIF programs. These involve gaming the system, perverse motivation, and turnover effects, among others. Evaluations might thus want to ask:

- Is there evidence that educators emphasized test preparation at the expense of in-depth instruction?
- Is there any evidence of breaches of test security or of falsifying of test scores?
- Do educators who never earn bonuses become demoralized and stop trying to improve performance?

- If incentives are based on individual teacher performance, has inter-teacher cooperation and collegiality decreased?
- If the program bases incentives on achieving schoolwide goals, do good teachers tend to leave schools that do not win bonuses to work in schools that do?

As is the case for all evaluation questions, when evaluators are able to deliver information on unintended consequences correctly to the LEA/SEA (See Section 5, Disseminating Evaluation Results), this information can have a significant impact on the program, as program staff can make adjustments to the inputs and associated activities to improve outputs and outcomes.

3 | Using Qualitative, Quantitative, and Mixed-Methods Approaches

This section addresses the appropriate application of qualitative, quantitative, and mixed-method approaches for measuring different aspects of the TIF program. The section focuses on how the specific evaluation question should determine which methodological approach to use in specific situations. This section encourages evaluators to use a balance of qualitative and quantitative approaches (mixed methods) to examine each of the inputs, activities, context, outputs, and short- and medium-term outcomes in a TIF program. Using a mixed-methods approach allows these methods to complement each other in ways that are beneficial to the evaluation audience. For example, if a TIF program is in the planning phases, an evaluator might need to use qualitative methods to explore the climate of a school during this planning phase to identify the perceptions of teachers/administrators toward performance pay and the particular program the district is implementing. Alternatively, once the district implements the program, the evaluator might need quantitative methods, such as an online survey, to measure the degree of stakeholder satisfaction.

Using a Qualitative Methods Approach

Qualitative methods are best suited for *exploring* relationships between variables/constructs. Evaluators use qualitative methods to develop a deep understanding of program inputs, activities, outputs, and outcomes and the relationships between each aspect of the logic model. Evaluators can also use qualitative methods to develop the logic model when it is not clear how to design a program.

TIF evaluations can use a wide range of qualitative tools, including:

- Proposals;
- Site visit reports;
- Participant observers' reports;
- Newspaper articles;
- Interview responses;
- Independent observers' field notes;
- Focus group transcripts; and
- Case studies.

Qualitative data can include a wide range of program variables, such as (a) beneficiaries' needs and wants, (b) how and why a program got started, (c) goals and plans, (d) schedules and budgets, (e) personnel and procedures, (f) equipment and facilities, (g) operations and expenditures, and (h) the intended and unintended outcomes.



Using a mixed-methods approach allows these methods to complement each other in ways that are beneficial to the evaluation audience.

Once a district collects the necessary qualitative data, the next step is data analysis. The Joint Committee on Standards for Educational Evaluation (1994) defines qualitative analysis as “the process of compiling, analyzing, and interpreting qualitative information about a program that will answer particular questions about that program” (p. 171). It is therefore important that analysis of qualitative information in an evaluation is appropriate and systematic so that evaluators can answer the research questions effectively.

For each set of qualitative information, the evaluators should choose an analytical procedure and plan for summarizing findings that are appropriate for addressing part or all of the evaluation’s questions and that suit the nature of the information to be analyzed. The three main categories of qualitative analysis strategies are *categorical*, *contextualizing*, and *thematic*.

Categorical strategies break down narrative data into smaller units and then rearrange those units to produce categories that facilitate a better understanding of the research question. Contextualizing strategies interpret narrative data within the context of the broader narrative, examining the connection among each of the narrative elements. A third analytical procedure for analyzing qualitative information is thematic analysis, which focuses on identifiable themes or patterns in the data. Thematic analysis requires the use of an explicit “code,” which may be a list of themes, a model that includes themes, or indicators. The theme is a pattern found within the data that describes and organizes the data, as well as helps interpret them. Evaluators can generate these themes either inductively from the set of data or deductively from a theory or prior research. Once the evaluator has established the themes and coded all of the data, the next step is to bring these themes together into a coherent explanation of the issue under analysis.

Using a Quantitative Methods Approach

Quantitative methods are best suited for *establishing* relationships between variables/constructs. TIF evaluations typically use a wide range of quantitative tools to gain a measurable understanding of program implementation and impact, including surveys, observations, and assessments. Within a quantitative evaluative framework, evaluators will operationally define aspects of the logic model for statistical analysis. Typical examples of constructs used in quantitative TIF evaluations include student achievement, teacher job satisfaction, principal leadership, teacher quality, professional learning community, parent satisfaction with school, student engagement, student school performance (grades), and measures of fidelity of implementation.

Statistical Analysis

Evaluators should start the quantitative statistical analysis process by exploring and gaining an understanding of the data set, identifying strengths and weaknesses in the data (including missing or miscoded data), making needed corrections, and discerning which available data can address the research questions. Evaluators should follow these steps with more systematic, often increasingly complex, analyses aimed at providing clear results and warranted interpretations. The evaluator should then reduce and synthesize the information to answer evaluation questions effectively. When synthesizing the information, the evaluator should provide tables, bar charts, and graphs so that stakeholders can understand the results.

Analyzing quasi-experimental designs is particularly challenging because nonrandom assignment of subjects to comparison groups introduces a host of difficulties in discerning whether observed between-group differences in outcomes were due to differences in treatments. Quantitative analysis in these situations requires careful model design: (a) rigorous

diagnostic analysis of the model and consequent results, (b) documentation of procedures used and the difficulties encountered in the analysis, and (c) followup of tests of main effects with tests of statistical interactions.

In any quantitative evaluation, it is important that evaluators are transparent about their methods and their analyses and that their calculations are defensible. As a rule, evaluators should document the procedures they used, state the assumptions these techniques required, report the extent to which the techniques met the assumptions, and justify their interpretations of the results of their analyses. In order to best maintain transparency and inform policymakers, evaluators should also take care in reporting potential weaknesses in the evaluation design or data analysis and discuss their possible influence on interpretations and conclusions. For example, if only a small number of schools are implementing performance incentives, the evaluation may not have the statistical power to infer causality (See Section 4, Evaluation Selection Framework).

Using a Mixed-Methods Approach

In a TIF program, qualitative and quantitative methods should inform each other and be used together. For example, if a TIF program is in the planning stages, evaluators can use qualitative methods such as interviews or focus groups to do a needs assessment to determine the preferred focus of the incentive plan. Once the program is in place, evaluators can use quantitative methods to determine what the impact of the incentive system is on student achievement. Ideally, evaluators will use qualitative and quantitative methods throughout the evaluation to measure different activities and their outcomes.

As discussed in the logic model/theory of action section, a systematic and comprehensive evaluation should answer more than just outcome questions.

Evaluations should also examine inputs, activities, context, outputs, and short- and medium-term outcomes. In order to achieve this balance, researchers should use both quantitative and qualitative approaches. When evaluators use both methodologies, the methods can complement each other in ways that are beneficial to the evaluation audience. While quantitative methods are standardized, efficient, and easily summarized, qualitative information can add depth and more ways to explore and understand quantitative findings.

A mixed-methods approach presents an alternative to solely quantitative and qualitative traditions by advocating the use of whatever methodological tools are required to answer the research questions under study. Tashakkori and Teddlie (2003) define mixed methods as “a type of research design in which qualitative and quantitative approaches are used in types of questions, research methods, data collection and analysis procedures, and/or inferences” (p.711). A mixed-methods approach to evaluation uses the strengths of both quantitative and qualitative methods to achieve systematic, comprehensive, and dependable findings (National Science Foundation, 1997). It is important that the designers of a mixed-methods approach select the appropriate combination of methods needed. A mixed-methods approach allows for both formative and summative assessment, which provides direction for program improvement and an assessment of program effectiveness over time. For examples of TIF evaluations using mixed-methods approaches, see Appendices 3: Chicago, 4: Ohio, 5: Philadelphia, and 6: Pittsburgh.

By using mixed-methods, evaluators can use triangulation to confirm research findings. Triangulation refers to the combinations and comparisons of multiple data sources, data collection and analysis procedures, research methods, investigators, and inferences that occur at the end of a study. As Webb, Campbell, Schwartz, and Sechrest (2000) have pointed out, “Once a proposition has been

confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. The most persuasive evidence comes through a triangulation of measurement processes” (p. 3).

A crucial activity of mixed-methods research is synthesizing the information from quantitative and qualitative analysis. Using the triangulation approach, evaluators use multiple information sources to support the validity of their conclusions and ultimately increase policymakers’ confidence in using results for decisionmaking (Shufflebeam and Shinkfield, 2007). One strong example of the benefits of synthesizing qualitative/quantitative data is the different strands of cost-benefit analysis.

Cost-Benefit and Cost-Effectiveness Analysis

Conducting an efficiency (cost-benefit or cost-effectiveness) analysis requires that evaluators gather both strong quantitative and qualitative data over the period of the evaluation. If evaluators use a mixed-methods approach to establish that an SEA, LEA, or school has implemented a program with fidelity and that the program has produced desired outcomes, it then becomes important for policymakers to ask two questions:

1. Is the program producing benefits sufficient to justify the costs?
2. How does the level of benefits the program is producing compare in cost to other interventions aimed at producing the same benefit?

Both methods are extremely important for program planners, policymakers, and taxpayers, as each group would like to know whether program investments are paying off in positive results that exceed those of similar programs (Kee, 1995; Tashakkori and Teddlie, 2003).

Cost-benefit analysis examines the relationship between program costs and outcomes, with both costs and outcomes expressed monetarily. It places a monetary value on program inputs and each identified outcome and determines the relationship between the monetary investment in a program and the extent of the positive or negative impact of the program. Through this process, cost-benefit analysis identifies a cost-benefit ratio and compares it to similar ratios for competing programs, providing information about comparative benefits and costs to policymakers. So long as monetary terms can describe the costs and benefits, this approach allows comparison among different projects with different goals. For example, the study of the Perry Preschool Program used cost-benefit analysis to determine the short-term and long-term benefits of a high-quality preschool program compared to other interventions (Barnett, 1996).

Cost-Effectiveness Analysis

Cost-effectiveness analysis examines the relationship between costs and outcomes in terms of the cost per unit of outcome achieved. Unlike cost-benefit analysis, both quality and quantity define cost-effectiveness or input (e.g., the number of teachers in a building and their qualifications). Evaluators gather this information through interviews, reports, or direct observations and then sum up the total cost of the ingredients. Typically, they divide this number by the number of students to get an average cost per student that they can measure against the effectiveness of the intervention. Evaluators can then make comparisons across interventions to inform decisionmaking (Levin and McEwan, 2001).

This section addressed the appropriate application of qualitative, quantitative, and mixed-method approaches for measuring different aspects of the TIF program. It encourages evaluators to use a balance of qualitative and quantitative approaches to examine each of the inputs, activities, context, outputs, and short- and medium-term outcomes within a TIF program.

4 | Evaluation Selection Framework

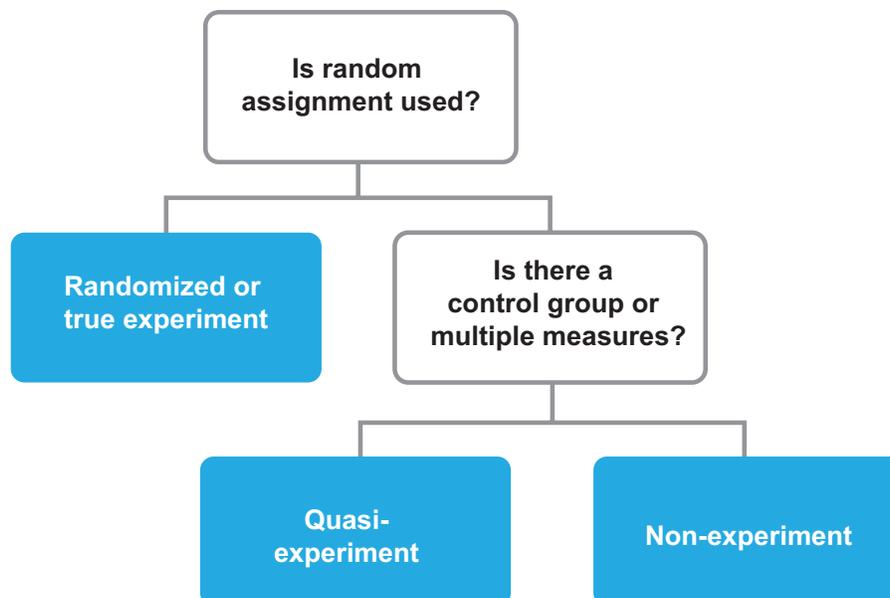
This section focuses on the appropriate selection framework for the evaluation, including experimental, quasi-experimental, and non-experimental designs (see Figure 4.1). More specifically, the section discusses the requirements for each framework and the strength of each at establishing causal relationships between an intervention (performance incentives for improved value-added scores) and an outcome (improved value-added scores). Moreover, the section addresses the importance of evaluators considering the type of program and available data when selecting the evaluation framework to ensure that the evaluation provides both a rigorous summative analysis of long-term outcomes (program impacts) and adequate information on inputs, outputs, and short-term outcomes for formative use.

Determining Rigorous Evaluation Selection Frameworks

If possible, evaluators should use strong experimental or quasi-experimental designs to answer the ultimate outcome questions, such as whether a treatment increases student performance. Experimental and quasi-experimental designs include randomized controlled trial experiments, matching studies, quasi-experiments, surveys using representative samples, and cohort/cross-sectional samples (Rossi, Freeman, and Wright, 1979; Teddlie and Tashakkori, 2008).

The following sections explain these methods in more detail. The goal of achieving internal and external validity should guide the selection of any of these approaches. It is crucial for evaluators to

Figure 4.1: Three major design options



design evaluations that try to establish a causal link between an intervention and an outcome. The strength of this linkage determines the level of internal validity. External validity is the degree to which conclusions about the evaluated intervention would hold for similar interventions in other places and times. (For more information on internal validity and criteria for meeting it, see Appendix 1.)

Designs for Answering Ultimate Outcome Questions

To answer the ultimate outcome questions, such as whether student achievement has increased, evaluators should use strong experimental or quasi-experimental designs. Some examples, noted above, include randomized controlled trial experiments, quasi-experiments, surveys using representative samples, and cohort/cross-sectional samples (Rossi et al., 1979; Teddlie and Tashakkori, 2008). The goal of achieving the greatest degree of internal and external validity should guide the selection from these approaches. This guidebook module focuses primarily on experimental and quasi-experimental designs, but a number of resources are available that discuss additional designs (Campbell and Stanley, 1963).

Treatment and Control Evaluation Designs

Randomized controlled experimental evaluation designs provide the strongest internal and external validity and, consequently, the most credible information about program outcomes. Not surprisingly, policymakers, stakeholders, and the general public often prefer these types of designs because they tend to provide the most convincing information about education programs (Nave, Miech, and Mosteller, 1998). Although experiments provide methodologically strong findings, conducting experiments can prove to be costly and difficult to implement (Podgursky and Springer, 2007). Logistically, it

can be costly and difficult to obtain consent from potential participants when there is no promise they will receive the program. From a political perspective, it can be difficult to convince schools and districts to use a randomized design because it requires them to withhold an intervention from a group of schools or students who may need or want it. In these cases, it is difficult to justify to “control” schools why they are not receiving the program. Although there are methods that may mitigate the push-back on districts attempting to implement randomized experiments, such as cross-over designs, where all schools or students ultimately receive the program, it still takes strong leadership and buy-in to implement this method successfully. Many TIF evaluations use both experimental and quasi-experimental designs to determine causal relationships between specified independent and dependent variables, such as incentivized professional development for principals and student value-added scores. These two design options vary, however, in their methods.

Randomized controlled trials are truly experimental in that they include a randomized treatment (intervention) and a control (no intervention) group. Quasi-experimental designs construct comparison groups using two major approaches—matching and statistically equating. Matching studies contrast participants and nonparticipants in programs for comparability in important respects. Statistically equating studies compare participants with nonparticipants while controlling statistically for measured differences between the two groups.

Randomized Treatment and Control Design

Many consider the randomized treatment and control (RTC) experiment to be the gold standard for assessing net impacts of interventions. The goal of these experiments is to isolate the effect of the evaluated intervention by ensuring that experimental and control groups are exactly comparable

except that one group received the intervention. This comparison between intervention and non-intervention requires, by definition, that only part of the targeted population receives the treatment (often referred to as partial-coverage programs). Once the evaluator determines the comparison groups, the logic of RTC is relatively simple. An RTC design compares outcomes of the experimental and control group participants by using statistical procedures to determine whether any observed differences are likely to be due to chance variations (Teddlie and Tashakkori, 2008). As mentioned before, due to the cost and difficulty in implementation, very few national or international evaluations of educational performance incentive programs have used the rigor of RTC design (Podgursky and Springer, 2007). For an example of a TIF evaluation using an RTC design, see Appendix 3: Chicago.

Random assignment of subjects to treatments is a core feature of the experimental design approach. Random assignment ensures that every experimental unit has an independent and equal chance of assignment to the experimental or control group. Consequently, the first step in conducting a randomized experiment is determining the units of analysis. The nature of the intervention and its targets will determine the choice of units of analysis in RTC. The randomly assigned experimental and

control units may be individual persons or intact groups of students, teachers, principals, or schools.⁶

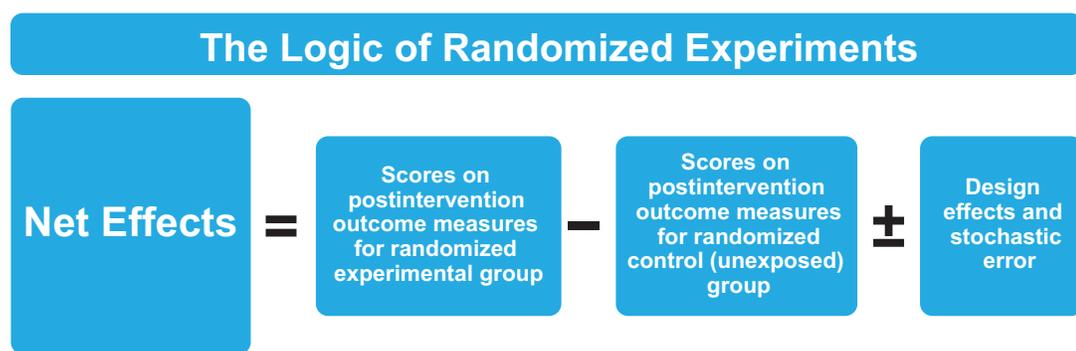
Randomly selecting individuals provides the researcher the greatest chance to detect a program effect. Randomly selecting 100 students in a school to participate and 100 as controls provides the researcher with greater statistical power than to select five classrooms to receive the program and five as controls. However, the integrity of randomized student selection is difficult to maintain; teachers and parents often treat control and participant students differently, thus contaminating the integrity of the program under investigation.

If the unit of selection in the RCT is classrooms or schools, then contamination is much less likely to occur. However, since randomizing from classrooms reduces the number of experimental units (a.k.a. sample size) to only a few, then the evaluation will be less likely to detect any treatment effect. In statistics, evaluators refer to this situation as having a low *power of analysis*.

Evaluators can follow a number of principles to increase the likelihood that the evaluation will be able to produce accurate results (a.k.a. statistical power). First, a formal power analysis should drive the number of students and/or sites planned

⁶ For an example of a TIF District using Randomized Treatment and Control Design see Appendix 3: Chicago Evaluation.

Figure 4.2: Logic of randomized treatment and control



for participation and control. Several programs to accomplish this are free, including Gpower (Buchner, Erdfelder, and Faul, 1996) and Optimal Design (Spybrook, Raudenbush, Congdon, and Martinez, 2009). To increase power to detect a program effect, the researcher could then match on relevant school, classroom, and/or student characteristics, increase the sample size, and collect time series data (Boruch, 2005).

The best experimental design occurs when groups are comparable across a number of dimensions, including composition (same units in terms of program-related and outcome-related characteristics), predisposition (equally disposed toward the project and equally likely to attain outcome), and experiences over the period of observation (same time-related, maturation, and interfering events). In practice, it is sufficient that the groups, as aggregates, are alike with respect to any characteristics that could be relevant to the intervention outcome.

Limitations of Randomized Experiments

While RTC experiments have earned the label of the gold standard for research design, designers must still weigh several limitations before choosing this methodology (Tashakkori and Teddlie 2003).

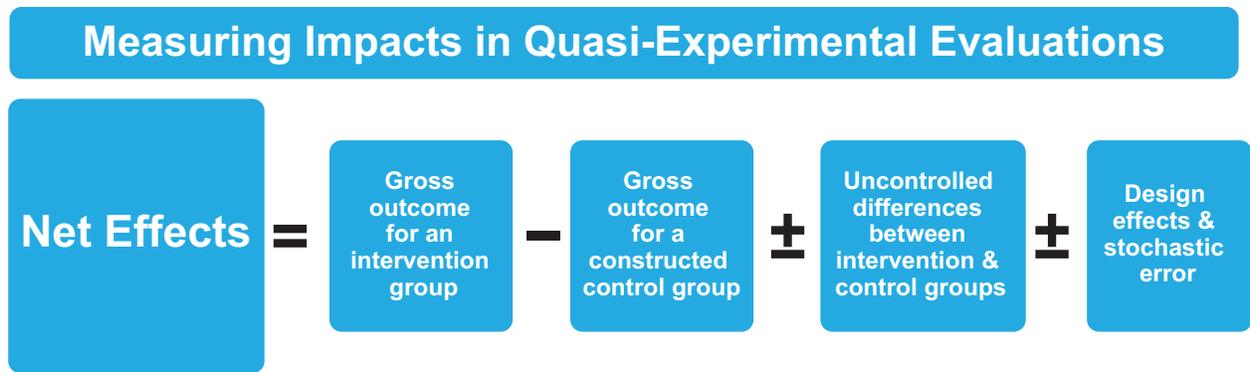
- **Ethics:** Stakeholders sometimes perceive randomization as unfair or unethical because of differences in the interventions given to experimental and control groups.
- **Early stages of program implementation:** RTC experiments may not be useful in the early stages of program implementation when interventions may change in ways the experiment does not allow.
- **Experimental intervention vs. intervention:** The way in which the experimental condition delivers the intervention may not resemble intervention delivery in the implemented program.

- **Cost and time required:** Experiments can be costly and time-consuming, especially large-scale, multi-site experiments.
- **Partial-coverage programs:** Randomized experimental designs are applicable only to partial-coverage programs in which there are sufficient numbers of nonparticipants from which to draw a control or comparison group.
- **Integrity of experiment:** Although randomly formed experimental and control groups are statistically equivalent at the start of an evaluation, nonrandom processes may threaten their equivalence as the experiment progresses.
- **Generalizability and external validity:** Because experiments require tight controls, evaluators may be limited in the degree to which they are able to generalize the evaluation results to other places, situations, and/or times (a.k.a. generalizability and external validity).

Quasi-Experimental Design Evaluations

Quasi-experimental designs are quantitative outcome designs that do not involve randomly assigned comparison groups. Evaluators usually select this type of evaluation because either the assignment to intervention and control condition is not within the evaluator's control or because of political, ethical, or other considerations that lead program staff, sponsors, or other powerful stakeholders to oppose randomization. While quasi-experiments do not involve random assignment of participants, they do require a well-defined and implemented treatment and, if there is a control group, that it be separate from the experimental treatment group. Quasi-experiments include *Ex Ante* (evaluators can choose how they will select the control group before the program is provided to an intervention group) and *Ex Post* (the evaluators develop the comparison group after the start

Figure 4.3: Logic of quasi-experimental design



of the intervention) designs. Evaluations use quasi-experiments to overcome threats to internal validity, and thus enhance their credibility when compared to studies that impose no controls on treatments and experimental subjects. Some argue that quasi-experimental designs have stronger external validity than true experiments because the latter often impose controls that would be hard to impose in the normal course of program delivery (Bracht and Glass, 1968).

Constructing Control and Comparison Groups in Quasi-Experimental Evaluations

The most common quasi-experimental designs involve constructing control or comparison groups in an attempt to approximate a randomized design. The major difference between quasi-experimental approaches is the way that the evaluator develops comparison and control groups to minimize the selection bias that results from the uncontrolled (i.e., nonrandom) assignment of targets to the experimental and comparison groups. Selection bias occurs when students, parents, or teachers have the opportunity to self-select into a program. In this situation, those who select into the program are likely different from those who opted not to participate. Perhaps they are more motivated, or

perhaps they have more involved parents. Typically, these differences are unmeasured and unknown, thus making it impossible to remove the bias from the analysis. Quasi-experiments provide evaluators with tools to, at least partially, address this. The two main quasi-experimental approaches are *matching* and *equating groups by statistical procedures* (for more on quasi-experimental design, see Campbell and Stanley (1963) and Cook and Campbell (1979)).

Constructing Control Groups by Matching

The matching process involves selecting units for control groups whose characteristics resemble the major relevant features of those units exposed to the program. For example, if evaluators choose a school as a target for the intervention, a matched control group would be one or more schools that have demographic profiles that mirror that of the participating school. An alternative is to select from within schools students who are similar to the participants. The options are thus either individual or aggregate matching.

In education, typically used individual controls and matching characteristics include age, sex, income, occupation, grade, free/reduced-price lunch eligibility, disability status, English Language Learner status, prior achievement, and race/ethnicity. At

larger levels, like classrooms or schools, aggregates or the individual characteristics can be used, in addition to class size, school size, teacher qualifications, and a multitude of other factors that could be relevant to a particular study.

One way to construct control groups by matching is by using a pre-post, nonequivalent comparison group design. This design is similar to the randomized experimental and control group, but in place of randomization, evaluators attempt to find a group as similar as possible to the one that will receive the new program by matching experimental and control group subjects. It logically follows that the pretest (such as prior achievement) is an important part of this design, particularly if it can help demonstrate equivalence of groups. For examples of TIF evaluations that are using quasi-experimental designs with matching, see Appendix 4: Ohio, and Appendix 5: Philadelphia.

Equating Groups by Statistical Procedures

To a large extent, evaluators have replaced or supplemented matching with the use of statistical controls to deal with selection bias or differences between groups. In this approach, evaluators collect information on the relevant variables for both the intervention and comparison groups and use statistical analyses to control for differences. Using a multivariate statistical model, meaning a model that includes multiple factors, evaluators can statistically control for individual and group-level differences. This model allows evaluators to make inferences about the remaining relationship between the interventions and the various measurable outcomes after accounting for the relationships between the other factors considered in the model (a.k.a. control variables) and the outcomes. An advantage to this approach is that the relationships among student

and school characteristics, program participation, and outcomes can be described using all students rather than a subset (i.e., sample) of students found to match on all control factors, which may therefore increase the statistical power to detect an effect.

Non-experimental Designs

Since non-experimental designs lack strength of causal inference and the internal/external validity of experimental and quasi-experimental designs, they are best for formative and implementation evaluation designs. A well-implemented non-experimental study can allow the evaluator to develop a deep understanding of the inputs, activities, context, outputs, and short-term/medium-term outcomes. Case studies and pattern matching are examples of how non-experimental designs can provide information about the implementation and effectiveness of TIF programs.

Case Study Evaluations

A case study evaluation's signature feature is an in-depth examination of the case in a detailed, descriptive report. The evaluator studies, analyzes, and describes the case as fully as possible. He or she examines the case's context, goals or aspirations, plans, resources, unique features, importance, noteworthy actions or operations, achievements, disappointments, needs and problems, and other topics. The evaluator reviews pertinent documents, conducts interviews with principal parties involved in the case or who are in a position to share insights about the case, and any other observable evidence. Using as many methods as necessary, the evaluator views the program in its different (and possibly opposing) dimensions as part of presenting a general characterization of the case. For an example of a TIF evaluation that is using a case study design, see Appendix 4: Ohio.



Overall, a pattern match illustrates a correspondence between the theoretical or conceptual expectation pattern and an observed or measured pattern.

Pattern Matching

Pattern matching is similar to case studies, in that there are no control groups. However, pattern-matching allows for more causal inference. While with case studies evaluators typically do not make specific predictions as to what they will find, if a program has a well-developed logic model, it may be possible for the evaluator to make specific predictions about what will be measured and when. If the evaluator verifies these predictions, he/she can make some causal inference that the program is having its intended effect. Overall, a pattern match illustrates a correspondence between the theoretical or conceptual expectation pattern and an observed or measured pattern. In program evaluation,

three pattern matches are important: the program pattern match that assesses program implementation; the measurement pattern match that assesses the validity of the measures; and the effect pattern match that assesses the causal hypothesis (Trochim, 1985). If the observed pattern across these areas matches the predicted pattern, the evaluator may be able to infer causation. The ability to infer causation through the development of a strong logic model makes this method preferred over case study designs.

5| Disseminating Evaluation Results

This section focuses on best practices for disseminating evaluation results to stakeholders. Evaluators must effectively communicate findings to stakeholders throughout the evaluation. When stakeholders understand formative and summative evaluation results, they are able to make programmatic decisions, such as whether they need to make improvements to and/or should continue the programs. The paragraphs that follow discuss strategies that evaluators can use to communicate evaluation results with stakeholders, such as establishing processes that encourage the use of evaluation findings, providing interim feedback, and agreeing on standards for the preparation and delivery of formative and summative reports.

Evaluators must effectively report their evaluation findings. In addition, evaluators should organize the findings to meet the needs of the various audiences, as well as provide stakeholders with the information that they need to make programmatic decisions. The evaluators' communication skills have a direct impact on whether the report will achieve its purpose of informing, educating, and convincing decisionmakers about ways to improve the program. Further, reports that do not appropriately report the methods and results of an evaluation can ruin the utility of the evaluation itself. The impact of an evaluation can extend beyond the particular evaluated program. For instance, the evaluation may also provide information that will inform implementation decisions in other contexts. The strategies articulated in the next four sections will assist evaluators in maximizing the impact of the evaluation results.

Arranging Conditions to Foster Use of Findings

A number of strategies are available to evaluators to increase the utility of evaluation results. First, evaluators should recognize the current makeup of the various audiences and stakeholders and take steps to involve these audiences on the front end to determine components of evaluation reporting. While much of the reporting schedule is determined in response to the RFP and prior to data collection and analysis, it is important that evaluators include stakeholders in these early conversations. These early conversations will not only serve broad engagement purposes, but also establish expectations about the format, style, and content of the final report (Stufflebeam and Shinkfield, 2007).



When stakeholders understand formative and summative evaluation results, they are able to make programmatic decisions.

Another strategy evaluators can use to improve how they communicate about the evaluation is to promote stakeholder buy-in by asking representatives from different interest groups to provide feedback on evaluation plans and instruments. Stakeholder groups may serve as key informants around how to navigate the contextual, programmatic, and political climate to maximize the utility of the evaluation. Ultimately, however, evaluators should maintain the authority to disagree with stakeholders when their input lacks logic and merit (Gangopadhyay, 2002). Section 6 in this guidebook explores this more fully.

Once evaluators and clients decide to proceed with an evaluation, they should negotiate a contract with strong provisions—budgetary and otherwise—for promoting effective use of evaluation findings. One strategy for involving stakeholders in the evaluation process is to develop an *evaluation review panel* that will provide feedback throughout the evaluation. The role of the panel is to review and provide feedback on draft evaluation designs, schedules, instruments, reports, and dissemination plans.

Providing Interim Feedback

A crucial part of communicating evaluation findings is interim reporting, which is typically part of the schedule for formative evaluation, but may also occur on an as-needed basis. The evaluator's response to the RFP should establish an expectation between the evaluator and the LEA/SEA for the amount of reporting, but the evaluators and the client must be flexible when unexpected events lead to the need to share information. For example, if problems occur with an incentive payout to principals or teachers, it is important for the district to share information about the problem so that the two parties can work together to establish the cause of the problem and its impact. Additionally,

evaluators should be open to ongoing interactions with stakeholders and be responsive to stakeholders' questions as they emerge, so that each group gets the information that it needs to make the program as effective as possible.

One way for evaluators to formalize productive interactions with stakeholders is to plan interim workshops with them (Gangopadhyay, 2002; Stufflebeam and Shinkfield, 2007). In this model, the evaluators send an interim report to the designated stakeholder group in advance of a feedback workshop and ask members to review findings and prepare questions in advance. During the workshop, stakeholders have opportunities to identify factual errors and ask pertinent questions about the evaluation. This process provides an opportunity for two-way communication and is an effective strategy for keeping interim feedback focused on program improvement needs. It also helps the client make immediate use of the findings for program improvement decisions.

Preparing and Delivering the Final Report

While the evaluator may present the final report (either formative or summative) in a number of ways, it is critical that the information it presents is well organized, aligned with the evaluation questions and expected evaluation process, and is clear, relevant, forceful, and convincing to stakeholders. The Joint Committee's Program Evaluation Standards (1994) emphasizes the importance of relevance to a variety of stakeholders by being comprehensive, clear, timely, and balanced. It is particularly important that evaluation reports are both comprehensive and reader friendly, a balance that often requires different versions of the report. In order to meet this balance between being comprehensive and user friendly, evaluation

reports should include an executive summary as well as the full report with findings and conclusions and should also include an appendix of evaluation methodology, tools, information collection, and data. Finally, in order for an evaluation to have its maximal impact for programmatic improvement and LEA/SEA decisionmaking, it is important that evaluators are sensitive and diplomatic about releasing evaluation information and balancing contractual and legal restraints with pressure from external audiences.

Presenting the Final Report

In addition to the report, evaluators should present evaluation findings verbally and visually to stakeholder groups. These presentations can range in intensity from simple PowerPoint presentations for district administrative staff to a series of workshops directed at teachers. If an evaluator wants the evaluation to make a difference and result in programmatic improvements, he/she must be committed to bringing the evaluation results to program staff. Evaluators cannot believe that simply writing their report will result in program staff following their recommendations and improving programs. Further, although the evaluation presentation is an opportunity to develop the knowledge of evaluation for district staff, the evaluator should be careful not to use too much technical jargon and instead rely on simple messaging strategies that address the main aspects of the evaluation.

Providing Follow-up Assistance to Increase Evaluation Impact

Providing a final report to stakeholders is not always enough to ensure that they act upon the findings in appropriate ways. Evaluators can provide follow-up assistance to stakeholders to increase the likelihood that programs will maximize evaluation results for program improvement. The evaluators can assist the client in determining ways to improve post-service reporting, such as identifying training needs of program staff, determining whether a new budget sufficiently addresses issues found in the program, increasing public understanding or acceptance of the program, or planning for a follow-up evaluation to address unidentified issues. The evaluator might continue to conduct workshops with relevant staff so that program staff can seriously consider and enact suggestions derived from formative and summative evaluation results.

6 | Managing TIF Program Evaluation Processes

This section guides TIF recipients through the process of developing evaluative management systems that promote the production of objective, high-quality evaluation. Though many of the challenges inherent in managing TIF program evaluation processes, such as deciding between internal or external evaluations, writing a Request for Proposals (RFP), selecting the evaluator, and developing a contract and scope of work are not specific to TIF, the complexity of TIF initiatives emphasizes the importance of TIF recipients making thoughtful decisions across all these processes. This section first discusses some of the challenges of conducting a useful and objective evaluation. Then it explores the conditions necessary for managing internal and external evaluations. The section concludes with a discussion of strategies TIF recipients can use to promote appropriate relationships with both internal and external evaluators and program staff, including strategies that TIF recipients can use for developing RFPs for evaluators, contracts, and budgets.

Challenges of Managing TIF Evaluations

Evaluators are in a powerful position because they or others can use their conclusions both to justify shutting down programs and firing staff, or alternatively, to expand programs. Therefore, evaluators must protect themselves from challenges both to their integrity and the integrity and quality of their evaluation. Since the value and usefulness of an evaluation requires objectivity, the evaluators must constantly demonstrate that they are not influenced by the client, their own beliefs, or current trends in performance-pay research. One challenge to the

objectivity of the evaluation is that program planners, developers, and implementation staff may attempt to influence the evaluators to make positive statements about the program. In this case, making negative attributions about the program could risk relationships with the program staff. This could result in accusations of bias toward the evaluators or the evaluation or program staff hiding the results or could even prevent the evaluator from evaluating programs in the future. It is important that evaluators take steps to ensure their objectivity and the results.

With TIF evaluations, the political dynamics have the potential to be even more complicated. TIF programs may have powerful individuals and groups both supporting and opposing them. TIF programs represent a paradigm shift in education; one from an entitlement human capital model to one that rewards teachers based on their productivity. With any paradigm shift, there are those who resist change, for whom a change of human capital management strategies in education could potentially usurp their power and control. Conversely, both the federal government and states have made significant investments with the hope that performance incentive programs can serve as an important mechanism for education reform in the United States. Either of these sides might challenge the validity of any evaluation (and the objectivity of its authors) that fails to support their initial views on the reform.

States and school districts often express general anxiety about the impact of the evaluation. This anxiety stems in part from political dynamics that may challenge the objectivity and integrity of TIF

evaluations. Donaldson, Gooler, and Scriven (2002) refer to fear and mistrust of evaluators by program staff as “evaluation anxiety.” Evaluation anxiety can be the result of previous bad experiences, a lack of experience with evaluators, a feeling of ownership over a program, or a fear of the potential consequences of negative findings. Table 6.1 summarizes the specific causes of evaluation anxiety at the individual and contextual levels, also referred to as the evaluation anxiety construct.

Table 6.1: Causes of evaluation anxiety

<p><u>Individual sources</u></p> <ul style="list-style-type: none"> • Lack of experience with program evaluation • Negative past experiences with program evaluation • Excessive ego involvement with program model • Excessive fear of negative consequences <p><u>Contextual sources</u></p> <ul style="list-style-type: none"> • Failure to highlight program accomplishments • Social norms • Role ambiguity <p><u>Interaction of individual sources and contextual sources</u></p>

The evaluation anxiety construct also addresses the various ways evaluation anxiety can manifest itself in the behaviors of stakeholders, which, in turn, could destroy an evaluation. Stakeholder resistance tactics might range from the more passive, such as hiding or minimizing program weaknesses, to the more aggressive, like accusing the evaluators of being biased or incompetent. Table 6.2 summarizes the various manifestations.

Table 6.2: Resistance to evaluation tactics

<ul style="list-style-type: none"> • Conflict—Accusing evaluators of hidden agendas • Withdrawal—Avoiding or refusing to work with evaluators • Resistance—Stalling, protesting, or failing to use evaluation results • Shame—Hiding weaknesses • Anger—Killing the messenger

Over the course of the evaluation, these tactics can wear down the evaluators into believing that a rigorous evaluation is pointless or impossible. If evaluators are unable to collect data because staff members have stopped cooperating, they have little opportunity to produce a valid or useful product. If staff members are openly hostile to evaluators, the evaluators might stop asking the tough questions or fail to document negative occurrences. The next section outlines many strategies for mitigating the risk of resistance stemming from evaluation anxiety.



*It is vital that grantees insulate those who conduct evaluations of TIF programs from the influence of others and from the **perception** of being influenced.*

Choosing the Type of Evaluator

TIF grant recipients must choose evaluators carefully. If grantees choose the wrong group or choose evaluators in an inappropriate manner, the integrity of the evaluation risks compromise. An evaluation that does not adequately insulate its staff and processes from those who have a stake in the program's outcome risks contamination. Generally, TIF recipients have three choices for types of evaluators to choose: internal, external, or a combination of both. The following sections discuss the implications of these.

Conducting the Evaluation Internally

Grantees should not take lightly the decision to design and implement a TIF evaluation internally. As discussed earlier in this guidebook, implementing and evaluating the TIF program can be politically sensitive to school districts and other stakeholders like teacher unions. Thus, it is vital that grantees insulate those who conduct evaluations of TIF programs from the influence of others and from the *perception* of being influenced. Both Stufflebeam (2002) and Volkov and King (2002) have outlined strategies for developing internal evaluation capacity that promote the successful implementation of internal evaluation, ensuring insulation from internal and external influences. In order to achieve this, TIF recipients should ask themselves the following questions when they choose an evaluation strategy:

1. Is the evaluation unit at a high enough organizational level to insulate it from inappropriate internal influences and enhance its influence on decisionmaking?
2. What parts of the evaluation does the evaluation team have the skills, leverage, and capacity to conduct well?

3. Is the district prepared to address challenges from external groups about the integrity of its evaluation?

Is the evaluation unit positioned at a high enough organizational level? This question assesses whether the evaluation unit can conduct a summative/ outcome evaluation of TIF. Generally, formative evaluations are less likely to induce evaluation anxiety than summative evaluations. If a TIF recipient decides to conduct a summative evaluation internally, it must position the evaluation unit at a high level in the organizational chart. Otherwise, there is a risk that the evaluators will fear retribution by program staff, which may prevent them from being honest in their evaluation. Alternatively, if the results of the evaluation are positive, positioning evaluation staff below program staff on the organization chart makes it likely that others will question the integrity of the evaluation. In this case, there may be an appearance that the evaluator has “colored” his/her characterization of the program either to please program staff or due to political pressure.

What parts of the evaluation does the evaluation team have the skills, leverage, and capacity to conduct well? This question speaks to the appropriateness of doing the formative or the summative evaluation internally. Stufflebeam (2002) lists the following expertise as necessary for an internal evaluation unit: field work, group process, interviewing, measurement, statistics, surveys, cost analysis, values analysis, policy analysis, public speaking, writing, editing, computing, communications technology, and project management (Stufflebeam, 2002). While not all these skills are necessary to conduct either a formative or summative TIF evaluation, the TIF recipient should understand its internal evaluation capacity to know what work is appropriate for it to do.

Is the district prepared to address challenges from external groups about the integrity of its evaluation?

Even if the evaluation unit is well insulated and highly skilled, the TIF recipient may still decide to conduct some or all of the evaluation externally. As mentioned elsewhere in this section, there is a difference between integrity and perceived integrity. Many people automatically view internal evaluations as biased, and given the political nature of TIF, it may be beneficial for some TIF recipients to excuse themselves from any part of the evaluation. Still, it is important to note that although using an external evaluator mitigates some of the danger that others will perceive the evaluation as biased, it does not necessarily mean that the evaluation is not free from bias. This guidebook explores this issue more in depth later.

Strategies for Conducting a Successful Internal Evaluation

Given the previous discussion about evaluation anxiety, internal evaluators must work intentionally to prevent the evaluation from turning negative. Donaldson, Gooler, and Scriven, 2002, outline several strategies for preventing or dealing with evaluation anxiety as it occurs. Six strategies are particularly important for TIF evaluations (Table 6.3).

Table 6.3: Strategies for addressing evaluation anxiety

1. Make sure resistance is not legitimate opposition to bad evaluation.
2. Determine program psychologic (term explained below).
3. Discuss why honesty with the evaluator is not disloyalty to the group.
4. Provide balanced continuous improvement feedback.
5. Allow stakeholders to discuss and affect the evaluation.
6. Distinguish the blame game from the program evaluation game.

Make sure resistance is not legitimate opposition to bad evaluation. Thus, always consider others' views of the evaluation first. As much as evaluators must overcome program staff feeling defensive about their programs, evaluators must overcome their own defensiveness about their evaluations. It is always possible that the criticisms are valid.

Determine program psychologic. Program psychologic refers to the individual fears and hopes that ride on the results of the evaluation. What weight do stakeholders place on the results of the evaluation? By recognizing these, the evaluators can develop their communication and collaboration strategies more intelligently, to be more sensitive to others and promote a more honest relationship.

Discuss why honesty with the evaluator is not disloyalty to the group. Education evaluation is a small world, and it is not always possible to completely disentangle personal relationships from professional ones. Given that evaluators and project staff often have long-standing relationships with one another, it is no surprise that project staff might view a negative evaluation as an act of betrayal. Still, for the most part, people are reasonable and understand the need for rigorous, objective evaluation results. Talking about this up front should help minimize the likelihood it will occur.

Providing balanced and continuous improvement feedback. Evaluators sometimes focus on the negative and ignore the positive. Although this is often born from a genuine desire to be helpful and demonstrate their usefulness, evaluators should outline both what is and is not working for a program. Further, evaluators should implement feedback systems that prevent conclusions from surprising stakeholders.

Distinguish the blame game from the program evaluation game. It is important that the tone of the evaluation not be accusatory. It is helpful to couch both positive and negative summative findings within contextually based explanations for why

the program did or did not work. The role of the evaluators is to identify the conditions that both promote and inhibit program success, not to blame individuals.

Strategies for Working with an External Evaluator

Generally, the process of working with an external evaluator involves three steps:

1. Developing an RFP
2. Selecting the evaluator
3. Defining the evaluator/stakeholder relationship.⁷

Navigating the RFP process

The fiscal agent (state, district, or not-for-profit organization) may issue an RFP to all potential evaluators or seek out specific evaluators with whom the agent has an established relationship or knows to have a reputation for excellence in a particular area. Some RFPs contain extremely detailed information on the project the grantee wants to evaluate and any previous evaluations that another evaluator may have performed, in addition to the specific requirements of the needed evaluation. Other RFPs are more general; the organization indicates that it wants the bidders to suggest necessary details and to exercise creativity.

Both highly specific and more general evaluation RFPs should indicate the evaluation's time line, main questions to be answered, needed information, the required reports, a recommended structure for proposals, the criteria for evaluating proposals, the deadline for submitting a proposal, references to relevant background materials, and the persons who can answer potential bidders' questions. In determining whether to respond to an RFP, it is important for evaluators to gauge the level of cooperation

they can reasonably expect to receive from program personnel; determine the accessibility of program materials; and glean the nature, quality, and availability of data from program records.

The following steps outline a basic process TIF recipients should use for selecting an external TIF evaluator. Most of what follows generalizes to other, non-TIF evaluations; however, TIF represents a unique set of projects, with various challenges common across TIF programs. Thus, the following process addresses these challenges. Regardless of the type of project, it is vital that the RFP process be objective, cost-effective, and result in an evaluation that will address both formative and summative project needs.

Step 1: Identify stakeholders and RFP committee participants

- Who is going to manage the evaluator's work, that is, at what organizational level will the evaluator report? It is important that this level be high enough to insulate the evaluator from potential pressure and influence from the program designers and implementers.
- Who should participate in the evaluator review process? Consider including a variety of representatives in the review process so that all stakeholder groups feel included. Doing so will increase the likelihood that stakeholder groups will be open to the evaluator, his/her activities, and his/her findings. Being inclusive and collaborative in the RFP and selection processes will result in a more successful evaluation.

Step 2: Define evaluation needs, that is, what questions is the evaluator to answer? With input from the identified representatives, does the project need summative evaluation support? Will the evaluators bid on providing formative evaluation information as well, or will the project be handling that

⁷Plans and Operations Checklist: http://www.wmich.edu/evalctr/archive_checklists/plans_operations.pdf

internally? Does the project need help developing a logic model and linking it to practice and the evaluation? Is the evaluator to provide technical assistance or at least present the result to various stakeholder groups, for example, school staff, district administrators, teacher unions, etc? It might be useful to put the evaluators in front of the dissemination process to prevent stakeholder groups from viewing the evaluation results as biased or influenced by the TIF 3 recipient.

Step 3: Identify adequate resources to fund the evaluation. The budget should be between 5 percent and 15 percent, depending on how great the need for formative evaluation support is.

Step 4: Develop the RFP: This should include:

- A list of the evaluation questions proposers need to answer. For TIF, at a minimum, proposers need to outline how they will answer the following questions:
 - o Did TIF improve student achievement by increasing teacher and principal effectiveness?
 - o How well did stakeholders understand the new compensation systems?
 - o How much “buy-in” did the TIF program have from the various stakeholders?
 - o How did TIF change the allocation of effective teachers across schools?
- Evaluators conducting a formative evaluation might also need to answer a myriad of additional questions, such as:
 - o What intermediate and short-term outcomes may lead to long-term outcomes such as improved student achievement and teacher attitudes toward the program, and how would you measure them?

- o How congruent is the espoused program logic model with the actual program in action?

- A requirement that the proposers summarize their experience with conducting school evaluations/TIF evaluations, and include specific work examples. It is also important that specific people be identified as responsible for the implementation of the evaluation. In larger evaluation firms, there is often a great deal of variability in the quality of work based on who is leading the evaluation. Grantees should be careful that the proposing organization is assigning staff to the project who have the necessary experience and skills. Further, the TIF program should ask for references and the right to follow up with any organizations that worked with the evaluator. Most larger evaluation firms have several positive clients to whom they typically refer potential clients. It is important to get information from these clients to find out how well things typically go.

Step 5: Assign points to the various pieces included in the RFP.

Step 6: Post and advertise the RFP. In addition to posting the RFP on the grantee website, TIF projects might consider posting it on message boards and the list serves for the American Evaluation Association and the American Educational Research Association. The process should allow potential applicants to ask questions. It is important that this process be as scripted as possible to prevent bias or the appearance of bias from seeping into the process.

Step 7: Before reviewing the proposals, design a review process. Questions to consider:

- Will the grantee be independently reviewing and scoring or reviewing as a group?

- Are there any individuals on the review panel who have a relationship with any of the proposers?
- Grantees might consider reviewing at least one proposal as a group to calibrate ratings.

Step 8: Check references and consider inviting top-rated proposers to present their evaluations.

Step 9: If the TIF grantee cannot make an obvious choice, the grantee should consider asking each finalist to make a final “best offer” for price and choose the one with the best price.

Agreeing on a Contract and Scope of Work

Once the RFP process has resulted in the selection of an evaluator, the grantee must then agree on a contract and scope of services. Stufflebeam (1999) developed a checklist as a tool to outline the specific components of evaluation contracts. If TIF recipients develop their evaluation contracts with this level of detail, the contracts will provide both parties with a clear understanding of their roles and expectations.

TIF grantees should not underestimate the importance of a sound evaluation. Grantees should appropriately negotiate contractual agreements that safeguard the evaluators’ ability to interact equitably and appropriately with all stakeholders and to ensure the study’s integrity. TIF recipients should negotiate a sound evaluation contract that helps set the conditions for disseminating evaluation findings effectively and provides a basis for settling disputes.⁸ Such contracts at a minimum should define

- the evaluator’s audience;
- the evaluation questions;
- the substance of interim and final reports;

⁸Negotiating Checklist: http://www.wmich.edu/evalctr/archive_checklists/negotiating.pdf



It takes careful planning to balance the scope of work for the evaluation with the funding, level of program cooperation, time line, and other essential resources allocated to the project.

- deadlines for submission;
- which audience segment will receive which reports;
- opportunities that stakeholders will have to contribute to the evaluation;
- authority for editing and disseminating reports;
- any provisions for pre-release review of reports;
- opportunities for program personnel to rebut reports; and
- provisions for reviewing and updating contractual agreements as needed.

Cronbach et al. have stated that, “deciding on a suitable level of expenditure is... one of the subtlest aspects of evaluation planning” (1980, p. 265). It takes careful planning to balance the scope of work for the evaluation with the funding, level of program cooperation, time line, and other essential resources allocated to the project.

The budget should align with the proposed evaluation design. The design should indicate the evaluation tasks, and an analysis of these tasks will indicate predictable costs. The evaluation design proposed through the RFP provides a forum for discussions and possible decisions, as LEAs and SEAs may be unaware of the extent of the information and costs an evaluation may produce. Items to consider in budgeting for an evaluation include personnel, materials, and the particular cost associated with each of the steps of the evaluation design. Stufflebeam has developed a useful checklist for constructing an evaluation budget (Stufflebeam and Shinkfield, 2007).⁹

Managing the Evaluation

Finally, the grantee should identify persons within the district to work with and supervise the work of the evaluator to avoid contamination of the evaluation at this point. If a stakeholder group, like program staff, manages the relationship with the evaluator, it is possible they will attempt to influence the findings of the evaluation. Through the effects of evaluation anxiety, they might do everything from block the evaluator from talking to certain persons or even refuse to accept the results of the evaluation.

Using Meta-Evaluation in Both Internal and External Evaluations

For both internal and external evaluations, we recommend that TIF recipients engage in a meta-evaluation process. Stufflebeam defines meta-evaluation as “the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation’s utility, feasibility, propriety, and accuracy and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide

the evaluation and publicly report its strengths and weaknesses” (Stufflebeam, 2001, p. 186). By hiring a separate evaluation group to conduct a meta-evaluation, grantees will further insulate the results of the summative TIF evaluation from influence and from skepticism. Meta-evaluations are a form of project management and thus free up internal staff from having to manage the day-to-day evaluation activities. Further, using meta-evaluation keeps the evaluator honest and prevents him/her from overcharging.

Finding a Balance

Between these two extremes of those who want to see TIF programs fail and those who think they are the answer to all the nation’s education programs lay the vast majority of individuals, who have not made up their mind yet about TIF programs. People generally are open-minded about the idea of TIF programs and wait to see the results of the TIF programs before they make a judgment.

Evaluators are the ones who will be determining the results, and in order to secure support for their findings, the evaluations must be valid, reliable, and free from undue influence. Regardless of whether the selected evaluators are internal or external, grantees can select and monitor them in a way that protects the integrity of the evaluation. In addition, it is just as important that the results of TIF evaluations be both valid and reliable. To that end, not all evaluation methodologies are equal. There are levels of rigor in both formative and summative evaluations that will determine the viability of the results of the evaluation. Hopefully, the use of this guidebook will improve both the rigor and integrity of TIF evaluations.

⁹See the resources at the Western Michigan University Evaluation Center. http://www.wmich.edu/evalctr/archive_checklists/evaluationbudgets.pdf

Appendix I | Internal and External Validity

Internal Validity

Internal validity, which measures the strength of causal relationships, is crucial in evaluation designs that try to establish a causal link between an intervention (such as teacher pay for value-added scores) and an outcome (improved value-added scores). The key question is whether outcomes or effects are the result of the program or intervention that the evaluator is studying or the result of other possible causes, such as contextual or demographic variables. Internal validity is only concerned with evidence that the specific program or intervention caused the observed outcome (Trochim, 2006).¹⁰ Research designs must meet certain criteria in order to establish internal validity. These include temporal precedence, co-variation of the cause and effect, and no-plausible-alternative explanation.

Temporal Precedence

To establish the criterion for temporal precedence, the evaluator must establish that the cause happened before the effect. This is often not difficult to do because most interventions occur prior to measurement of effects.

Co-variation of the Cause and Effect

The criterion for co-variation of the cause and effect requires that the evaluator establish a relationship between the intervention and the outcomes. In other words, evaluators meet the criterion if they observe that whenever the intervention is present, the outcome is also present *and* that the intervention is not present when the outcome is not present.

¹⁰<http://www.socialresearchmethods.net/kb/intval.php>

Sometimes there is an interest in establishing a continuous relationship—that is, whether different amounts of the intervention lead to different amounts of the outcomes (e.g., bigger recruitment incentives lead to higher quality teachers). Evaluators meet the criterion for co-variation of the cause and effect so long as they establish a comparison group that does not receive the intervention.

No-Plausible-Alternative Explanation

The criterion for no plausible alternative explanation requires that the evaluator establish that the intervention is causing the effect instead of a “plausible alternative.”¹¹ Typically, evaluators measure the particular outcome under analysis (e.g., student achievement) before implementing an intervention in order to establish a baseline. A year later, evaluators measure student achievement again to assess whether student performance has improved. Yet, even if student achievement goes up, a number of plausible alternative explanations unrelated to the program, such as changes in the student population, might cause the observed increase in the outcome measure. The no-plausible-alternative explanation criterion illustrates the importance of a research design that identifies each of the threats to internal validity and shows whether there truly is a causal relationship between the intervention and outcome variables.

¹¹For more on single-group threats, multiple-group threats, social threats, see Campbell and Stanley (1963) and Trochim (2006) (<http://www.socialresearchmethods.net/kb/intsing.php>).

External Validity

Researchers define external validity as the ability to generalize the findings from the research design to similar situations in the general unstudied population. In other words, it is the degree to which conclusions about the evaluated intervention would hold for similar interventions in other places and times. Two ways to make a study generalizable are sampling and proximal similarity.

In the sampling approach, the evaluators draw a representative sample from the target population and then generalize to the entire population to assess the likely impact of the program. In order to draw the most representative sample, evaluators should look at as many sources of data as are available.

In the proximal similarity approach, the evaluators' charge is to consider different generalizability contexts and assess which contexts are most like the study and which are least like it (Campbell 2002). By establishing similar contexts according to a number of factors (e.g., persons, places, or times), the evaluator can establish the degree to which the two contexts are similar. From this proximal framework, the evaluator can make greater generalizations to persons, places, or times that are more similar. The threats to external validity are the degree to which the evaluators are wrong about the similarity between these factors. Within this proximal approach, external validity can be improved through thorough descriptions of the way in which contextual factors are the same and different.

Appendix 2 | Joint Committee Standards

The Joint Committee, developed in 1975, is a professional association located at the University of Iowa concerned with the quality of evaluation. In 1981 the Joint Committee published its initial set of standards for evaluations of educational programs, projects, and materials. In order to stay current, the Joint Committee engages in an ongoing process of revising these standards.

While the following will provide a general description of each of these areas, the Joint Committee Standards include a tremendous amount of resources for information collection activities, including developing instrument blueprints, constructing response items, drafting and pilot-testing instruments, performing item analysis, performing reliability and validity studies, selecting appropriate samples of respondents, controlling information collection conditions, verifying obtained data, and keeping collected information secure.¹²

Information Scope and Selection

Joint Committee Standard: “Information collected should be broadly selected to address pertinent questions about the program and be responsive to the needs and interests of clients and other specified stakeholders” (Joint Committee on Standards for Educational Evaluation, 1994, p. 37).

The evaluators should collect information that has sufficient scope to address the audience’s most important information needs by obtaining information on all the important variables (e.g., beneficiaries’ needs and participation, program goals and

assumptions, program design and implementation, program costs and outcomes, and positive and negative side effects). The reality, however, is that evaluators need to be selective in deciding which information to collect because it is often not possible to meet all the information needs of the stakeholders.

Rights of Human Subjects

Joint Committee Standard: “Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects” (Joint Committee on Standards for Educational Evaluation, 1994, p. 93).

Since evaluators gather information from and pertaining to a wide range of persons associated with the subject program (program beneficiaries, staff, administrators, policymakers, community members, and others) they must make provisions for adhering to all applicable rights of those who are included in the evaluation. An effective way of upholding rights of human subjects is to vet an evaluation design through the appropriate institutional review board.

Program Documentation

Joint Committee Standard: “The program being evaluated should be described and documented clearly and accurately, so that the program is clearly identified” (Joint Committee on Standards for Educational Evaluation, 1994, p. 127).

This standard emphasizes that when the evaluators release the evaluation report, the range of readers who will read it should know about how the evaluators conceived and implemented the program. The original program proposal provides insufficient

¹² Evaluation Standards: Program Evaluation Standards: <http://www.jcsee.org/program-evaluation-standards/program-evaluation-standards-statements>

information because implementation may have been very different from what was proposed. For example, a TIF program may experience political pressures that led to altering the incentive structure for teachers so that the grant awarded a higher proportion of compensation for teacher participation in professional development than for increasing student achievement scores.

It is important that evaluators document the implementation in detail not only for program improvement, but also for others considering adopting and implementing the program or some modification of it. Also, if a program fails, the program funders will need to have information on program expenditures, staffing, and operations in order to determine the reasons for failure. Finally, researchers of incentive programs who are interested in the program's effects need detailed information about the programs' actual operations so they can relate parts of the program to its outcomes.

Context Analysis

Joint Committee Standard: "The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified" (Joint Committee on Standards for Educational Evaluation, 1994, p. 133).

Contextual factors have a significant impact on program design and operation as well as on what the program achieves; therefore, evaluators need to collect a considerable amount of this information. Evaluators should consider a program's geographical location, political and social realities, the economic health of the relevant community, program-related needs, how and why the program started, related legislation, and related state and national influences. This information is particularly important for formative evaluation, to assist stakeholders in

taking account of local circumstances and also to determine how the program is meeting the needs of targeted constituents. Summative evaluations rely on context to assist stakeholders in understanding reasons outside of the program that could have led to its success or failure. Potential examples of sources for contextual information are demographic information, economic data, or relevant legislation.

Defensible Information Sources

Joint Committee Standard: "The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed" (Joint Committee on Standards for Educational Evaluation, 1994, p. 141).

Evaluators should rely on multiple sources of information to provide cross-checks and triangulate findings and achieve a greater sense of accuracy. Evaluators can draw from a wide range of sampling techniques (simple random, stratified random, purposive, snowball) to improve validity and reliability. Since evaluation is a time-bound process, evaluators should strive for representativeness and transparency, while also being straightforward about limitations of their information sources. The Joint Committee on Standards for Educational Evaluation (1994) advises evaluators to "document, justify, and report their sources of information, the criteria and methods used to select them, the means used to obtain information from them, and any unique and biasing features of the obtained information" (p.141). One way evaluators can report this information is through a technical appendix that includes information sources, the information collection process, and the instruments used to collect the information (Teddle and Tashakkori, 2008).

Valid Information

Joint Committee Standard: “The information gathering procedures should be chosen or developed and then implemented, so that they will assure that the interpretation arrived at is valid for the intended use” (Joint Committee on Standards for Educational Evaluation, 1994, p. 145).

Researchers define validity concerns as the soundness and defensibility of inferences or conclusions drawn from the information-gathering processes and products. Evaluators can use information-gathering products and associated processes such as results of interviews, observations, document reviews, focus groups, and administration of rating scales. Evaluators should choose and employ processes that produce information that is relevant to study questions, reliable, and sufficient in scope and depth to answer all of the evaluation’s questions (Tashakkori and Teddlie, 2003, pp. 564, 563).

Reliable Information

Joint Committee Standard: “The information gathering procedures should be chosen or developed and then implemented, so that they will assure that the information obtained is sufficiently reliable for the intended use” (Joint Committee on Standards for Educational Evaluation, 1994, p. 153).

An evaluative conclusion cannot be valid if it is based on unreliable information. Information is unreliable to the extent that it contains unexplained contradictions and inconsistencies or if evaluators would obtain different answers under subsequent but similar information collection conditions, absent a known intervention. Information is reliable when its consistency is evident; it is free of internal

contradictions; and, when repeated, information collection episodes would, as expected, yield the same answers.

One gauges the reliability of information by examining its amount and types of variation, including desired or explainable variation and unwanted variation. Most information-gathering procedures give information with some amount of internal disagreement, or if applied repeatedly, give at least slightly different answers between settings, groups, and different times of collection. Depending on the nature of the evaluation, evaluators can be concerned about different forms of reliability: stability, equivalence, and internal consistency (Trochim, 2006).¹³ It is important that evaluators report weaknesses in obtained information and warn readers to be cautious in the use of their findings.

Systematic Information

Joint Committee Standard: “The information collected, processed, and reported in an evaluation should be systematically reviewed, and any errors found should be corrected” (Joint Committee on Standards for Educational Evaluation, 1994, p. 159).

Systematic information control is an information management process to ensure that evaluators regularly and carefully check, make as error-free as possible, and keep secure the evaluation’s information. Evaluators must avoid numerous errors, including mistakes in collecting, scoring, coding, recording, organizing, filing, releasing, analyzing, and reporting information. Evaluators should institute safeguards to prevent all such mistakes.

¹³<http://www.socialresearchmethods.net/kb/relytypes.php>

Appendix 3 | Chicago Evaluation—Randomized Control Trial with Quasi-Experimental Matching

Program Characteristics¹⁵

Based on the national Teacher Advancement Program (TAP) model, Chicago TAP aims to improve schools by raising teacher quality through monetary incentives tied to student achievement, among other strategies. Chicago TAP seeks to support and develop high-quality teaching by offering sustainable opportunities for career advancement and ongoing school-based professional development, by insisting on instructionally focused accountability, and by providing performance pay.

Methods

An external evaluator, Mathematica, designed an evaluation using experimental (random-assignment) and quasi-experimental (propensity score matching) methods to estimate the impact of Chicago TAP. Sixteen elementary schools in the Chicago Public Schools (CPS) voluntarily applied for TAP and successfully completed the TAP selection process. Mathematica randomly assigned eight TAP schools to a treatment group that began implementing TAP in 2007-2008 and eight to a control group that delayed implementation until 2008-2009. Mathematica complemented the experimental analysis with a comparison sample of 18 additional schools by matching them according to size, average teacher experience, and student demographics. They repeated this design in 2009, randomly assigning another 16 schools to the sample, half of which were scheduled to implement TAP in 2009-2010 and the other half to delay implementation until 2010-2011.

Data

Mathematica administered a teacher questionnaire in spring 2008, interviewed principals in fall 2008, and obtained CPS student test score files and teacher administrative records covering all years since 2006-2007. Mathematica is continuing the process of collecting follow-up data.

¹⁵ Chicago TIF Grant: <http://www.cccr.ed.gov/initiatives/prof les/pdfs/Chicago.pdf>

Appendix 4 | Ohio Evaluation—Quasi-Experimental and Comparative Case Study

Program Characteristics¹⁶

The Ohio Teacher Incentive Fund (OTIF) program includes schools in Cincinnati, Columbus, Toledo, and Cleveland. The goal of OTIF is to improve student achievement and increase the number of effective teachers assigned to disadvantaged and minority students in hard-to-staff areas through a number of pay-for-performance policies. The program designer built OTIF based on existing models, including the Teacher Advancement Program (TAP) in Cincinnati and Columbus and the Toledo Review and Alternative Compensation System (TRACS). Cleveland has implemented a new program called Promoting Educator Advancement in Cleveland (PEAC).

Methods

To measure changes in student achievement, the evaluator (Westat) is using a quasi-experimental design in two districts in which not all schools are participating. Westat's design uses treatment and comparison schools that it matches on student achievement, socioeconomic status, minority enrollment, and size. In two other districts in which all schools are participating, the study uses a time-series design that compares trend data from all schools that were collected before and after program implementation. Westat is using teacher surveys and developing case studies to examine program impact.

Data

Westat is collecting data from all four large urban districts participating in OTIF. Westat administers annual teacher surveys to all teachers in participating OTIF schools in two nonsaturation districts and to a random sample of teachers in two districts using a saturation model. The survey includes questions on support for and knowledge of the program, perceived changes in working conditions, attitudes about financial incentives, and related issues. Case studies of 12 schools across the four districts include interviews with teachers, principals, and district/union staff. Westat is also collecting financial allocations and expenditures data annually.

¹⁶ Ohio TIF Grant: <http://www.cccr.ed.gov/initiatives/profles/pdfs/Ohio.pdf>

Appendix 5 | Philadelphia Evaluation— Quasi-Experimental and Comparative Case Study

Program Characteristics¹⁷

The purpose of the Philadelphia TIF project (Promoting Excellence in Philadelphia Schools, or PEPS), is to pilot a performance-based staff development and compensation system that provides teacher and principal incentives tied directly to student achievement growth and classroom evaluation. The evaluator (Temple University) is examining both the implementation and impact of a national model, the Teacher Advancement Program (TAP), in a cluster of public charter schools in the School District of Philadelphia. The evaluation seeks to understand whether components of the program are being implemented with fidelity, and if so, how. The evaluation also seeks to understand the prevalence of local adaptation, effects on teacher quality, and linkages to student achievement. The evaluation design tests the theory of action for differentiated incentives within a school and determines whether incentive models spark whole school change.

Methods

For the evaluation, Temple is conducting a dual research design: quasi-experimental and comparative case study. In the quasi-experimental portion of the evaluation, Temple matched intervention sites with comparable schools in the district. Factors for matching were student demographics, school size, grade span, and poverty. Temple is also using an abbreviated time series design to determine

within-school changes in student achievement and teacher retention. Temple is using the case study design to determine how the context affects implementation and outcomes in each of the participating schools.

Data

Temple is using the following sources of data to evaluate the program:

- interviews with school-based leadership teams (i.e., principal, master teacher, mentor teacher) and key program staff;
- school observations of cluster group meetings and leadership team professional development sessions;
- school context variables (drawn from observational data, along with data collected for the U.S. Department of Education's Common Core of Data);
- high-stakes standardized student achievement test scores from state and district assessment programs; and
- researcher-developed surveys of staff perceptions.

Additionally, Temple is using reviews of participating schools and surveys of teachers conducted by the national TAP program as supplemental data.

¹⁷ Philadelphia TIF Grant: <http://www.cccr.ed.gov/initiatives/profles/pdfs/Philadelphia.pdf>

Appendix 6 | Pittsburgh Evaluation—Quasi-Experimental Design with Implementation Analysis

Program Characteristics¹⁸

The Pittsburgh TIF Program (Pittsburgh Principal Incentive Program, or PPIP) is intended to promote instructional leadership by providing incentives and assistance to help principals improve their practices. PPIP includes two components:

1. an annual salary increment based on a rubric that principals' supervisors administer, which measures practices in seven areas; and
2. an annual bonus based primarily on student achievement growth.

Methods

An external evaluator, RAND, has designed the evaluation with the intention of helping the district identify and address strengths and weaknesses in the measures and in the overall program. The evaluation uses a mixed-methods research design that includes

1. documentation of program activities, including steps taken to ensure high-quality performance measures;
2. analysis of principals' buy-in, participation in professional development, and leadership practices;
3. analysis of the validity and reliability of the measures;
4. correlational and regression analyses to explore relationships among the survey measures, scores on the rubric and bonus measures, and school and principal characteristics; and

5. examination of trends in principal performance and student achievement over time.

One of the primary methodological challenges that RAND faces is the absence of a comparison group, because the program was adopted district wide.

Data

For the implementation evaluation, RAND is using documents developed by the Pittsburgh Public Schools and its partners, including meeting minutes, professional development materials, program documents (e.g., the evaluation rubric), and email communication with district staff. The evaluation of outcomes uses data sources such as principal and teacher questionnaires, interviews with principals and district staff, school site visits, principal scores on the evaluation rubric, and student test scores and demographic data.

¹⁸ Pittsburgh TIF Grant: <http://www.cccr.ed.gov/initiatives/profles/pdfs/Pittsburgh.pdf>

Appendix 7 | Power/Causality/Feasibility Analysis

	Power	Causality	Feasibility
RTC at the student level	High	High	Medium
RTC at the school level	Moderate	High	High
Regression discontinuity with program test score cutoff	Low	High	High
Year fixed effects, Covariates	Low	Low	High

1. Randomized Controlled Trial at the student level.

Power: Highest.

Causality: High—gives best causal evidence if implemented with fidelity.

Feasibility: Medium—will require strong commitment from both district and school officials. Low to medium fidelity is possible.

2. Randomized Controlled Trial at the school level.

Power: Moderate—randomization of 16 schools will not produce good power.

Causality: High—gives best causal evidence if implemented with fidelity.

Feasibility: High—only requires district buy-in.

3. Regression Discontinuity with program test score cutoff.

Power: Low—less than random assignment.

Causality: High—strong causal evidence.

Feasibility: High—requires schools to follow simple assignment rule.

4. Year fixed effects, covariates

Power: Low—due to fixed effect.

Causality: Low—year fixed effects and student characteristics help control for selection.

Feasibility: High—important to not assign student based on unobservables.

Bibliography

- Barnett, W. S. (1996). *Lives in the balance: Age-27 benefit-cost analysis of the High/Scope Perry Preschool Program*. (Monographs of the High/Scope Educational Research Foundation, 14). Ypsilanti, MI: High/Scope Press.
- Boruch, R. F. (2005). *Place randomized trials: Experimental tests of public policy*. Thousand Oaks, CA: Sage Publications, Inc.
- Bracht, G. H., and Glass, G.V. (1968). The external validity of experiments. *American Educational Research Journal*, 5(4), 437–474.
- Buchner, A., Erdfelder, E., and Faul, F., (1996). G Power: A General Power Analysis Program. *Behavioral Research Methods*, 28(1), 1-11.
- Campbell, D. T., and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin Company.
- Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Florence, KY: Cengage Learning.
- Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco: Jossey-Bass.
- Donaldson, S. I., Gooler, L. E., and Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation*, 23, 261–273.
- Frechtling, J. A. (2007). *Logic modeling methods in program evaluation*. San Francisco: Jossey-Bass.
- Gangopadhyay (2002). *Making evaluation meaningful to all education stakeholders*. Retrieved Dec 2010, from http://www.wmich.edu/evalctr/archive_checklists/makin-gevalmeaningful.pdf
- Goldhaber, D. (2009). Politics of teacher pay reform. In M. G. Springer (Ed.), *Performance incentives: Their growing impact on American K-12 education*. (pg. 120-132). Washington, DC: Brookings Institution Press.
- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs*. (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Kee, J. E. (1995). Benefit and cost analysis in program evaluation. In J. S. Wholey, H. P. Hatry, and K. E. Newcomer (Eds.): *Handbook of practical program evaluation* (pp. 456–488). San Francisco: Jossey-Bass.
- National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation, and Communication. (1997 August). *User-friendly handbook for mixed method evaluations*. J. Frechtling and L. Sharp (Eds.). Washington, DC: Author.
- Levin, H. M., and McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications*. (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Nave, B., Miech, E. J., and Mosteller, F. (2000). A rare design: The role of field trials in evaluating school practices. In D. L. Stufflebeam, G. F. Madaus, and T. Kellaghan, (Eds.). *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, second edition. Boston: Kluwer Academic Press.
- Podgursky, M. J., and Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, 26(4), 909–949.
- Rivkin, S.G., Hanushek, E. A., and Kain J. F. (2002). *New evidence about Brown v. Board of Education: The complex effects of school racial composition on achievement*. NBER Working Papers 8741, National Bureau of Economic Research, Inc.
- Rossi, P. H., Freeman, H. E., and Wright, S. R. (1979). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage Publications, Inc.
- Shadish, W. R., Cook, T. D., Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Spybrook, J., Raudenbush, S. W., Congdon, R., and Martinez, A. (2009). *Optimal design for longitudinal and multilevel research: Documentation for the optimal design*. Software V.2.0.
- Stufflebeam, D. (1999). *Evaluation contracts checklist*. Accessed September 9, 2010, from www.wmich.edu/evalctr/checklists/
- Stufflebeam, D. (2001). The metaevaluation imperative. *American Journal of Evaluation*, 22, 183–209.

- Stufflebeam, D. L. (2002). *Institutionalizing evaluation checklist*. http://www.wmich.edu/evalctr/archive_checklists/institutionalizingeval.pdf
- Stufflebeam, D. L., and Shinkfield, A. J. (2007). *Evaluation theory, models, & applications*. San Francisco: Jossey-Bass.
- Tashakkori, A., and Teddlie, C. (Eds.) (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage Publications, Inc.
- Teddlie, C. B., and Tashakkori, A. (Eds.). (2008). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage Publications, Inc.
- Trochim, W. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review*, 9(5), 575–604.
- Trochim, W. M. (2006). *The research methods knowledge base*, 2nd Edition. Accessed May 10, 2011, from <http://www.socialresearchmethods.net/kb/>
- Volkov, B., and King, J. (2002). *A checklist for building organizational capacity*. Accessed October 20, 2010, from http://www.wmich.edu/evalctr/archive_checklists/ecb.pdf
- Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (2000). *Unobtrusive measures*. (Revised ed.) Thousand Oaks, CA: Sage Publications, Inc.

Program Evaluation for the Design and Implementation of Performance-Based Compensation Systems

Peter Witham, University of Wisconsin, Madison

Curtis Jones, University of Wisconsin, Madison

Anthony Milanowski, Westat

Christopher Thorn, University of Wisconsin, Madison

Steven Kimball, University of Wisconsin, Madison

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the suggested citation is: Witham, P., Jones, C., Milanowski, A., Thorn, C. and Kimball, S. Program Evaluation for the Design and Implementation of Performance-Based Compensation Systems. Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington D.C., 2011.

The Center for Educator Compensation Reform (CECR) was awarded to Westat — in partnership with Learning Point Associates, Synergy Enterprises Inc., Vanderbilt University, and the University of Wisconsin — by the U.S. Department of Education in October 2006.

The primary purpose of CECR is to support Teacher Incentive Fund (TIF) grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources. CECR also is charged with raising national awareness of alternative and effective strategies for educator compensation through a newsletter, a Web-based clearinghouse, and other outreach activities.

This work was originally produced in whole or in part by the Center for Educator Compensation Reform (CECR) with funds from the U.S. Department of Education under contract number ED-06-CO-0110. The content does not necessarily reflect the position or policy of CECR or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by CECR or the federal government.



Center for
Educator Compensation
Reform

Allison Henderson, Director

Phone: 888-202-1513

E-mail: cecr@westat.com

