Teacher & Leadership Programs

tif TEACHER INCENTIVE FUND + tsl TEACHER & SCHOOL I.EADER INCENTIVE PROGRAM

Evaluating Programs for Strengthening Teaching and Leadership

Teacher & Leadership Programs

tif TEACHER INCENTIVE FUND + tsl TEACHER & SCHOOL LEADER INCENTIVE PROGRAM

Evaluating Programs for Strengthing Teaching and Leadership

2016

Revised by:

Anthony Milanowski
Westat

Matthew Finster
Westat

Based on the document originally authored by:

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

The U.S. Department of Education (ED) expects all Teacher Incentive Fund (TIF) grantees to conduct an evaluation of their programs. Experience with earlier rounds of TIF grants has shown that evaluations can provide valuable information for managing and improving TIF-supported activities, as well as evidence that these activities have had a positive impact that can help grantees make the case for sustaining the work after the grant has ended. In addition, ED is interested in learning how TIF grants have improved educator effectiveness, student achievement, and equity of access to teachers for students from low-income families and minority students across and within schools and districts. This guidebook provides strategies and resources to address the many complexities and challenges involved in evaluating a TIF program, from conceptualizing the evaluation questions to reporting results. 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 design for assessing impacts (Section 4). The guidebook also describes 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 Expected Connection Between TIF Activities and Desired Outcomes?

This section provides both program designers and evaluators with resources for the conceptualization phase of the evaluation. Specifically, the section articulates how TIF program designers can develop or refine the logic model ED required them to develop for the grant application to further specify the expected causal relationships between TIF-funded activities and their intended program goals.

This section provides examples of common theories of action at work in TIF programs and strategies for making these theories concrete. Clarifying the theory of action and developing the logic model allows the TIF evaluator to construct appropriate evaluation questions to establish whether the program is accomplishing its goals.

Section 2: Developing Evaluation Questions

Section 2 provides strategies for using the logic model and theories of action to create targeted, formative and summative evaluation questions. It shows how grantees can structure the evaluation questions around the inputs, activities, outputs, and short- or medium-term outcomes the logic model identifies for use in both formative (periodic) and 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 of Program Impacts

This section focuses on helping evaluators determine an appropriate evaluation design for examining the impacts of TIF-funded activities on important outcomes such as educator effectiveness, student achievement, and equitable access to effective teachers. It also discusses experimental, quasi-experimental, and nonexperimental designs. This section discusses the requirements and strengths of each design and 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 describes best practices for disseminating 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: Managing TIF Program Evaluation Processes

This final section guides TIF grantees through the process of developing systems that promote objective, high-quality evaluations. It addresses the importance of choosing the right person to conduct the evaluation and outlines the decisions that a project director should make in choosing who will conduct the formative and summative evaluations of the TIF grant. 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.

1 Logic Models and Theories of Change

As part of their applications, ED required TIF grantees to develop a logic model that identifies the key components of their TIF programs and describes the relationships among the key components and the intended outcomes. Logic models help program designers and evaluators to conceptualize how they expect program activities to work together to influence desired outcomes. The value of a logic model is its clear representation of overall structure of the program and the connections among program inputs, activities, outputs, and outcomes (described below). The term “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 program outputs). By visually depicting the causal chain, the logic model helps an evaluator think about how to construct evaluation questions that will answer whether the program is “doing what it is supposed to do.”

Logic Models

This section reviews the concept of a logic model, provides an example of a logic model for a TIF grant, and shows how grantees can use theories of change or action as the basis for developing more detailed logic models covering particular sets of TIF-supported activities. We discuss the use of logic models to develop evaluation questions further in Section 2.

The main parts of a logic model are:

Program inputs: the resources the program uses to start and sustain it. TIF project inputs might include the funds provided by the federal government, the project staff hired with these funds, and the support of important stakeholders.

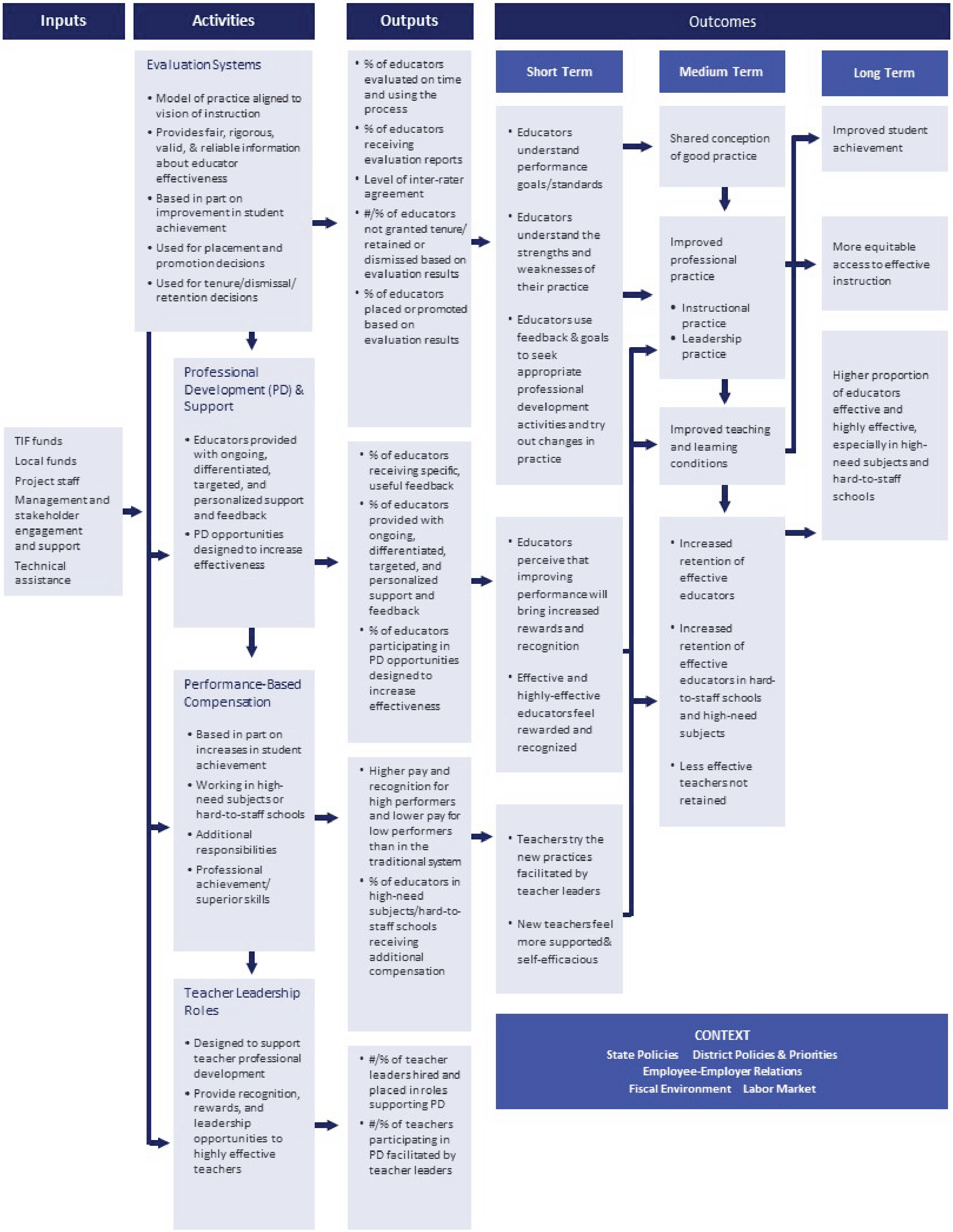
Program activities: the tasks and operations that program staff and others engage in to achieve program goals. Central to most TIF programs are activities such as evaluating educator effectiveness, selecting teachers for leadership roles, providing professional development opportunities based on evaluation results, and providing performance-based compensation to effective educators.

Program outputs: the direct results or products of program activities. According to Frechtling (2007), outputs are the most immediate indicators that the activities are actually being implemented, a fundamental prerequisite for the causal links represented in the logic to be made. Evaluators should identify at least one output for each activity, in order to guide the collection of evidence that the project has implemented the activity successfully. In most TIF projects, outputs will include the number or proportion of educators evaluated and provided with feedback, participation on professional development activities, receipt of performance-based compensation, and the proportion of tenure/placement or retention decisions projects made using evaluation results.

Program outcomes: the results of program activities, such as increased student achievement. Often, logic models distinguish between short- and medium-term outcomes, such as changes in teaching practices and retention, and ultimate outcomes, such as improved student achievement and equity in access to effective instruction.

Logic models may also include important contextual factors that can influence how strongly the inputs and activities affect the outputs and outcomes. For many TIF grantees, 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 state accountability systems.

The figure below shows an example of a logic model for a TIF grant. This logic model is based on the required aspects of a TIF proposal, as described in the Request for Applications.



Inputs

Activities

Outputs

Outcomes

Short Term

Medium Term

Long Term

TIF funds

Local funds

Project staff

Management and stakeholder engagement and support

Technical assistance

Evaluation Systems

- Model of practice aligned to vision of instruction
- Provides fair, rigorous, valid, & reliable information about educator effectiveness
- Based in part on improvement in student achievement
- Used for placement and promotion decisions
- Used for tenure/dismissal/ retention decisions

Professional Development (PD) & Support

- Educators provided with ongoing, differentiated, targeted, and personalized support and feedback
- PD opportunities designed to increase effectiveness

Performance-Based Compensation

- Based in part on increases in student achievement
- Working in high-need subjects or hard-to-staff schools
- Additional responsibilities
- Professional achievement/ superior skills

Teacher Leadership Roles

- Designed to support teacher professional development
- Provide recognition, rewards, and leadership opportunities to highly effective teachers

- % of educators evaluated on time and using the process
- % of educators receiving evaluation reports
- Level of inter-rater agreement
- #/% of educators not granted tenure/ retained or dismissed based on evaluation results
- % of educators placed or promoted based on evaluation results

- % of educators receiving specific, useful feedback
- % of educators provided with ongoing, differentiated, targeted, and personalized support and feedback
- % of educators participating in PD opportunities designed to increase effectiveness

- Higher pay and recognition for high performers and lower pay for low performers than in the traditional system
- % of educators in high-need subjects/hard-to-staff schools receiving additional compensation

- #/% of teacher leaders hired and placed in roles supporting PD
- #/% of teachers participating in PD facilitated by teacher leaders

- Educators understand performance goals/standards
- Educators understand the strengths and weaknesses of their practice
- Educators use feedback & goals to seek appropriate professional development activities and try out changes in practice

- Educators perceive that improving performance will bring increased rewards and recognition
- Effective and highly-effective educators feel rewarded and recognized

- Teachers try the new practices facilitated by teacher leaders
- New teachers feel more supported & self-efficacious

Shared conception of good practice

Improved professional practice

- Instructional practice
- Leadership practice

Improved teaching and learning conditions

- Increased retention of effective educators
- Increased retention of effective educators in hard-to-staff schools and high-need subjects
- Less effective teachers not retained

Improved student achievement

More equitable access to effective instruction

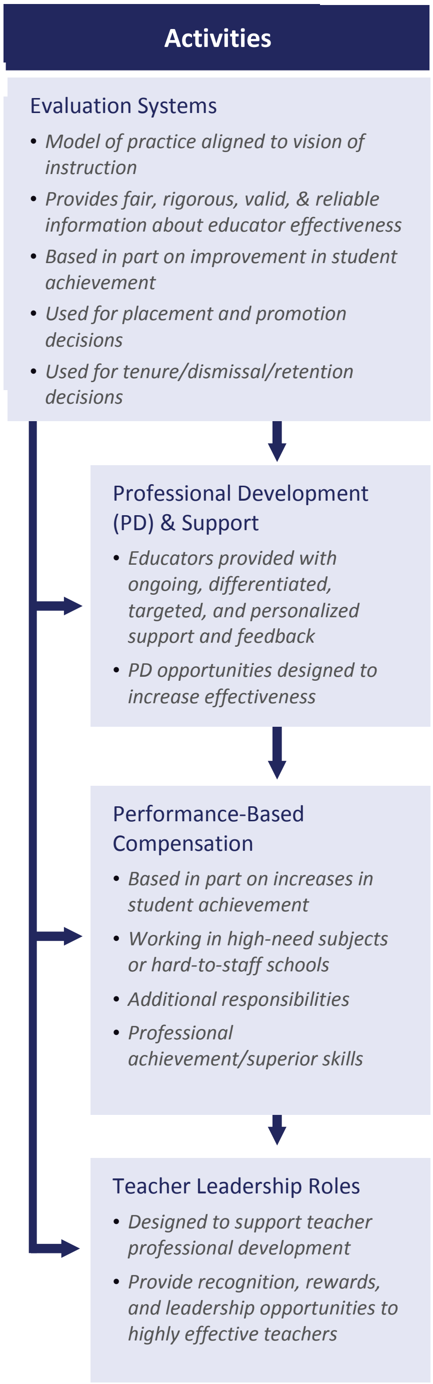
Higher proportion of educators effective and highly effective, especially in high-need subjects and hard-to-staff schools

CONTEXT

State Policies District Policies & Priorities

Employee-Employer Relations

Fiscal Environment Labor Market



Activities

Evaluation Systems

- *Model of practice aligned to vision of instruction*
- *Provides fair, rigorous, valid, & reliable information about educator effectiveness*
- *Based in part on improvement in student achievement*
- *Used for placement and promotion decisions*
- *Used for tenure/dismissal/retention decisions*

Professional Development (PD) & Support

- *Educators provided with ongoing, differentiated, targeted, and personalized support and feedback*
- *PD opportunities designed to increase effectiveness*

Performance-Based Compensation

- *Based in part on increases in student achievement*
- *Working in high-need subjects or hard-to-staff schools*
- *Additional responsibilities*
- *Professional achievement/superior skills*

Teacher Leadership Roles

- *Designed to support teacher professional development*
- *Provide recognition, rewards, and leadership opportunities to highly effective teachers*

Inputs

In the example logic model, the inputs include the funding the TIF grant provided, additional local funds fulfilling the TIF match requirements, staff assigned to the project, the support and engagement of local education agency (LEA) management and stakeholders (e.g., school board, teacher union), and technical assistance available from the TIF Technical Assistance Network and/or other providers supported with TIF funds.

Activities

Once these resources are in place, program administrators can carry out a set of activities that begin with program design and continue with communication of important program features (e.g., standards of educator performance, number and timing of observations, and how compensation will link to performance). TIF projects should promote the development and refinement of human capital management systems centering on educator evaluation and support. Thus, the activities represented in the example logic model include evaluation of teachers and principals based in part on demonstrated improvement in student academic achievement and the use of the evaluation process to provide educators with ongoing, differentiated targets and personalized support and feedback, supported by professional development opportunities designed to increase effectiveness. Grantees should also differentiate compensation based in part on measurable increases in student academic achievement and may reward responsibilities and effectiveness in hard-to-staff schools or high-need subject areas or assignment of additional responsibilities or job functions, such as teacher leadership roles and evidence of professional achievement and mastery of content knowledge and superior teaching and leadership skills. In addition, TIF expects grantees to use evaluation results for other human capital management decisions, including recruitment and hiring as well as placement and promotion, tenure/dismissal, and retention. These are the foundation of the TIF human capital management system the TIF grant program defined. Beyond these basics, TIF grantees can develop and deploy a variety of activities to improve the educator work force.

Outputs

Program outputs (measurable products of program activities) include the number or proportion of educators evaluated using the evaluation system, the number or percentage receiving specific feedback and ongoing support and participating in the professional development opportunities designed to increase effectiveness, and the number or proportion eligible for and receiving performance-based compensation and/or additional compensation for working in high-need subjects or hard-to-staff schools. Grantees using TIF to support teacher leadership programs might also include the number of teacher leader positions identified and filled and how often or for how many colleagues are teacher leaders carrying out roles like leading professional development, serving as mentors, or modeling effective instruction.

Outputs

- *% of educators evaluated on time and using the process*
- *% of educators receiving evaluation reports*
- *Level of inter-rater agreement*
- *#/% of educators not granted tenure/retained or dismissed based on evaluation results*
- *% of educators placed or promoted based on evaluation results*

- *% of educators receiving specific, useful feedback*
- *% of educators provided with ongoing, differentiated, targeted, and personalized support and feedback*
- *% of educators participating in PD opportunities designed to increase effectiveness*

- *Higher pay and recognition for high performers and lower pay for low performers than in the traditional system*
- *% of educators in high-need subjects/ hard-to-staff schools receiving additional compensation*

- *#/% of teacher leaders hired and placed in roles supporting PD*
- *#/% of teachers participating in PD facilitated by teacher leaders*

The example logic model divides outcomes into short-, medium-, and long-term categories. This division is useful because TIF activities may take a considerable amount of time to influence some of the intended outcomes, and many have precursors or preconditions that must be in place before they can achieve ultimate outcomes. Distinguishing between short-, medium-, and long-term outcomes allows the representation of causal chains that lead from activities to the outcomes the programs are ultimately intended to achieve. While there is no hard and fast definition of short, medium, and long term, it can be useful to think of short-term outcomes as the perceptions, attitudes, and behaviors of the educators most affected by the activities and outputs. Typically, TIF-supported activities affect longer-term outcomes through changes in educator behavior, which is influenced by educators' perceptions and attitudes. Thus, evaluators may want to discover how the activities and output are influencing these immediate outcomes.

Short Term

- *Educators understand performance goals/standards*
- *Educators understand the strengths and weaknesses of their practice*
- *Educators use feedback & goals to seek appropriate professional development activities and try out changes in practice*

- *Educators perceive that improving performance will bring increased rewards and recognition*
- *Effective and highly effective educators feel rewarded and recognized*

- *Teachers try the new practices facilitated by teacher leaders*
- *New teachers feel more supported & self-efficacious*

Short-Term Outcomes

In the example, short-term perceptual/attitudinal outcomes include educator understanding of the performance goals or standards and of the strengths and weaknesses of their practice, perceptions of greater reward and recognition for effective and especially highly effective educators, and perceptions of support and self-efficacy of new teachers supported by teacher leaders serving as mentors.

Short-term behavioral outcomes include making use of the evaluation feedback to select professional development and making practice changes and participation in professional development activities facilitated by teacher leaders.

Medium Term

Shared conception of good practice

Improved professional practice

- *Instructional practice*
- *Leadership practice*

Improved teaching and learning conditions

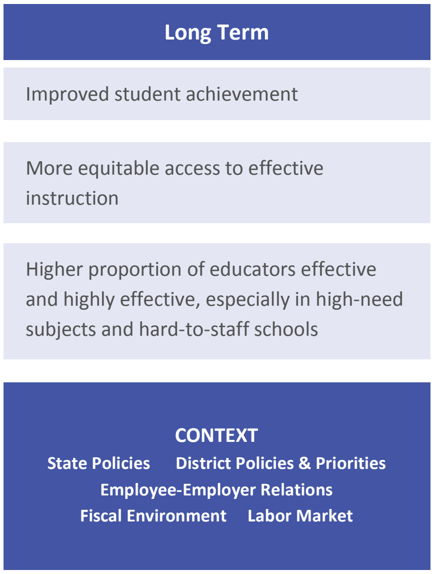
- *Increased retention of effective educators*
- *Increased retention of effective educators in hard-to-staff schools and high-need subjects*
- *Less effective teachers not retained*

Medium-Term Outcomes

Medium-term outcomes often are the ongoing patterns of behavior that result from perceptions, attitudes, and decisions made by program participants. In the example logic model, one of these outcomes is the development of a shared conception of good practice, reflecting the vision of instruction underlying the evaluation and support system. By defining what it means to be a good educator, this shared conception in turn reinforces changes in practice that are consistent with the performance standards through peer pressure and educators’ desire to fit in with the school or district culture. This outcome may represent a culture change in some schools, if there had been no coherent instructional vision or if new performance standards are more rigorous, and prior evaluation practices were lax.

The second medium-term outcome is improved professional practice. In particular, educators’ use of evaluation feedback and professional development builds skills needed to improve practice; recognition and rewards support motivation to improve practice; and a shared conception of good practice provides social reinforcement for acting in accordance with the vision of instruction underlying the evaluation and support activities. Improved practice includes improved instruction by teachers and improved support for teaching and learning by school leaders.

The logic model example also postulates that TIF activities will affect teacher retention, improving retention of effective educators and decreasing retention of poor performers. In turn, this outcome leads to an increase in the average level of professional practice and a higher proportion of effective and highly effective educators, especially in high-need subjects and hard-to-staff schools. The latter outcome overlaps the medium- and long-term categories because these effects should continue over the life of the TIF program.


Long Term
Improved student achievement
More equitable access to effective instruction
Higher proportion of educators effective and highly effective, especially in high-need subjects and hard-to-staff schools
CONTEXT
State Policies
District Policies & Priorities
Employee-Employer Relations
Fiscal Environment
Labor Market

Long-Term Outcomes

The two long-term, "bottom line" outcomes are improved student achievement, due to improved instructional practice and teaching and learning conditions, and more equitable student access to effective instruction, especially in high-need subjects and hard-to-staff schools.

Contextual Factors

Contextual factors can limit the impacts of TIF activities, or they can augment them. The logic model thus should include important contextual factors. The example logic model shows five categories of contextual factors likely to influence TIF projects:

- State policies such as accountability systems and educator evaluation regulations. For example, state testing systems can provide data needed for measuring student achievement growth, but also may not cover all grades and subjects or may change in the middle of a grant. Accountability systems may not be consistent with district performance standards, especially for school leaders, causing potential confusion about priorities and reporting requirements.

- District policies and priorities, including other instructional initiatives. For example, a new curriculum could reinforce the effect of the TIF program, if it aligns 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 TIF program if it did not align with the TIF evaluation and support system and diverted leaders' attention from implementing TIF activities as intended.

- Employee-employer relations, including collective bargaining agreements (where applicable) that may govern evaluation practices, teacher assignment, and salary schedules.

- Fiscal conditions, and especially changes in fiscal capacity, can influence whether educators believe that performance-based compensation will be viable as well as whether grantees can sustain TIF activities throughout and after the grant period.

- The external labor market is another type of contextual factor. Shortages or surpluses of high-quality recent graduates of teacher preparation programs can influence how selective districts or schools can be in hiring new teachers, or how much emphasis they put on removing ineffective ones. This dynamic in turn affects the demand for induction support and mentor workloads, as well as how quickly removing ineffective teachers can improve overall teaching quality.

During the first four rounds of TIF, all of these factors influenced the evolution of TIF activities and the success of particular grantees in implementing the scope of work for which they originally received the grant. Including important contextual features in the logic model reminds evaluators to be on the lookout for complicating or countervailing effects.

A logic model is useful for both formative and summative evaluation purposes. It identifies key activities, outputs, and short-term outcomes that are expected if grantees implemented the program components as intended. A logic model helps evaluators understand the program designers' intent, by identifying the links in the causal chain from inputs to ultimate outcomes. From a formative perspective, most program administrators will want to know as soon as possible if the intended activities are not taking place as planned and expected outputs produced, so they can take corrective action as needed. The logic model can also help structure an assessment of implementation fidelity, the degree to which a grantee implemented the program components as intended. Fidelity is important to program designers, who want to see the program implemented as designed. The degree of fidelity is also important for summative evaluation of program impacts because only if a grantee implements a program as designed can stakeholders know whether the program leads to the outcomes it was designed to produce. An evaluation structured by the logic model could also provide evidence for attributing long-term outcomes to the TIF activities, even in the absence of a comparison group. If the intended long-term outcomes occur, they are more likely to be due to TIF activities if the evaluation shows that grantees performed program activities as intended and produced the outputs intended and that these outputs produced the short-term outcomes expected followed by the medium- and long-term outcomes.

What is a theory of change?

Theories of change (also sometimes called theories of action) map out what needs to happen in order for the activities or strategies chosen to affect the long-term outcomes we wish to achieve. They can be expressed in a series of if… then… statements that postulate that if an activity of strategy is implemented, then we will see certain immediate outcomes, that in turn are pre-conditions (if's) for the longer-term outcomes we desire. Theories of change thus consist of causal links between activities and outcomes. It is often useful to represent these links in causal diagrams, with boxes or circles representing activities and outcomes and arrows representing causal links. Diagramming the theory of change for specific activities supported by your TIF project makes causal links and presuppositions clear and enables you to construct more detailed logic models and identify evaluation

Developing More Detailed Logic Models for Program Components

While the overall TIF logic model provides a useful overview of what grantees are to evaluate, it is often hard to represent all of the important causal links among components of the program that grantees need to make for them to achieve their goals. In particular, it can be difficult to show the causal connections between specific activities and each short-, medium-, and long-term outcome. This difficulty may inhibit developing evaluation questions that address all of the factors that contribute to the success of the program. Frechtling (2007) described how to develop subordinate logic models that focus on parts of the overall program and thus allow specification of relationships and outcomes in more detail. However, to develop these more detailed models, it is often useful to consider the theory of change or theory of action for how grantees expect the program components to cause the desired outcomes and to develop a diagram that shows the causal links.

While most program staff and evaluators have an intuitive notion of how they intended the program to work, at times their ideas differ, and they overlook important intermediate steps. Explicitly diagramming the theory of change will put program staff and evaluators in a better position to develop more detailed logic models for specific parts of their TIF programs and bring to the surface assumptions about the change process that may not be obvious. Developing an explicit theory of change also prompts evaluators to ask about each link in the causal chain from input to outcome. This enables examining program effectiveness at each link, thereby providing formative information about the program's impact and implementation fidelity. Understanding the theory of change or action may require evaluators to interview program designers and administrators and read program documentation. Because program administrators may not have a completely worked out model in mind, it may also be helpful to review prior research and theory. The goal is to establish how grantees expect program activities to affect outcomes and to specify the important causal links that evaluation questions should address.

Because a theory of change diagram does not typically depict outputs, nor is a strong distinction made between short-, medium-, and long-term outcomes, as in a logic model, it is sometimes easier for grantees to show a more detailed picture of the conditions that they expect to lead to the ultimate outcomes and to show the web of interconnected causal links between the outcomes.

The rest of this section considers theories of change for three components of a TIF-supported human capital management system in detail, to show how grantees can depict these theories, and suggests causal links and intermediate outcomes evaluations could address. Grantees could use the three diagrams shown to develop subordinate logic models by adding program outputs or as supplements to the overall TIF logic model.

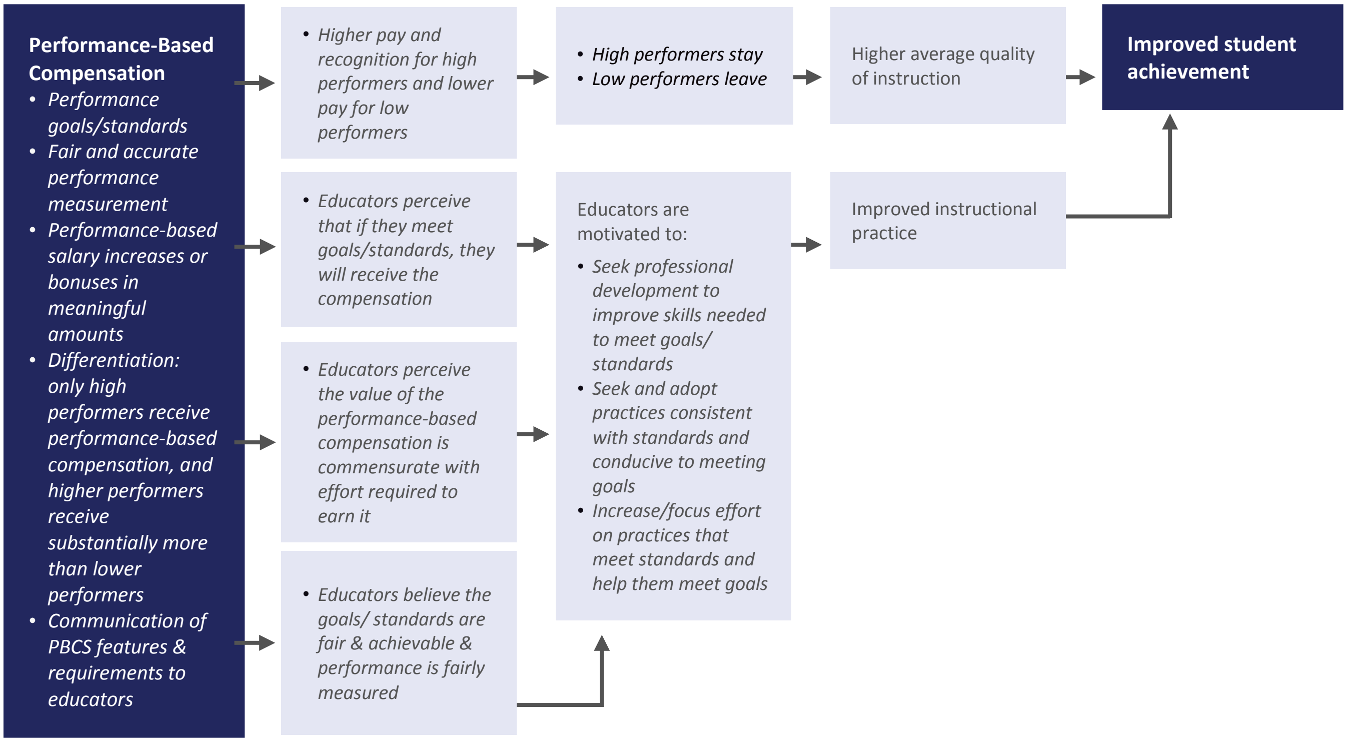
Theories of Change

Theory of Change for Performance-Based Compensation

The basic logic behind performance-based compensation systems (PBCS) is that linking compensation to measures of performance will influence educator behaviors that in turn influence student achievement. Examples of behavior changes could include increasing effort toward goals the system rewards (e.g., student achievement growth), searching for and adopting instructional practices thought to improve student achievement, gaining additional knowledge and skills from professional development offerings, or moving to schools where performance incentives are available. The figure below illustrates a generic theory of change for performance-based compensation.

The specific program elements, or "active ingredients," of performance-based compensation include performance goals or standards that define what levels of performance are required for educators to receive financial rewards and recognition, a measurement system to establish whether performance has met the goals or standards, and salary increases or bonuses of sufficient size to be meaningful and are differentiated so that higher performers would receive substantially more than lower performers. Grantees communicate these features of the PBCS to educators in ways that help them understand the performance requirements, how performance is measured, and the rewards and recognition available.

Performance-based compensation leads to improved **student achievement**.



Performance-Based Compensation

- *Performance goals/standards*
- *Fair and accurate performance measurement*
- *Performance-based salary increases or bonuses in meaningful amounts*
- *Differentiation: only high performers receive performance-based compensation, and higher performers receive substantially more than lower performers*
- *Communication of PBCS features & requirements to educators*

- *Higher pay and recognition for high performers and lower pay for low performers*

- *Educators perceive that if they meet goals/standards, they will receive the compensation*

- *Educators perceive the value of the performance-based compensation is commensurate with effort required to earn it*

- *Educators believe the goals/ standards are fair & achievable & performance is fairly measured*

- *High performers stay*
- *Low performers leave*

Educators are motivated to:

- *Seek professional development to improve skills needed to meet goals/ standards*
- *Seek and adopt practices consistent with standards and conducive to meeting goals*
- *Increase/focus effort on practices that meet standards and help them meet goals*

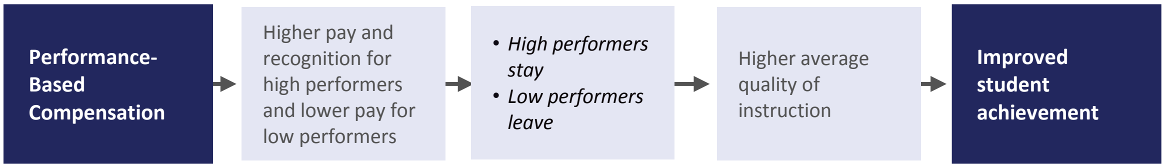
Higher average quality of instruction

Improved instructional practice

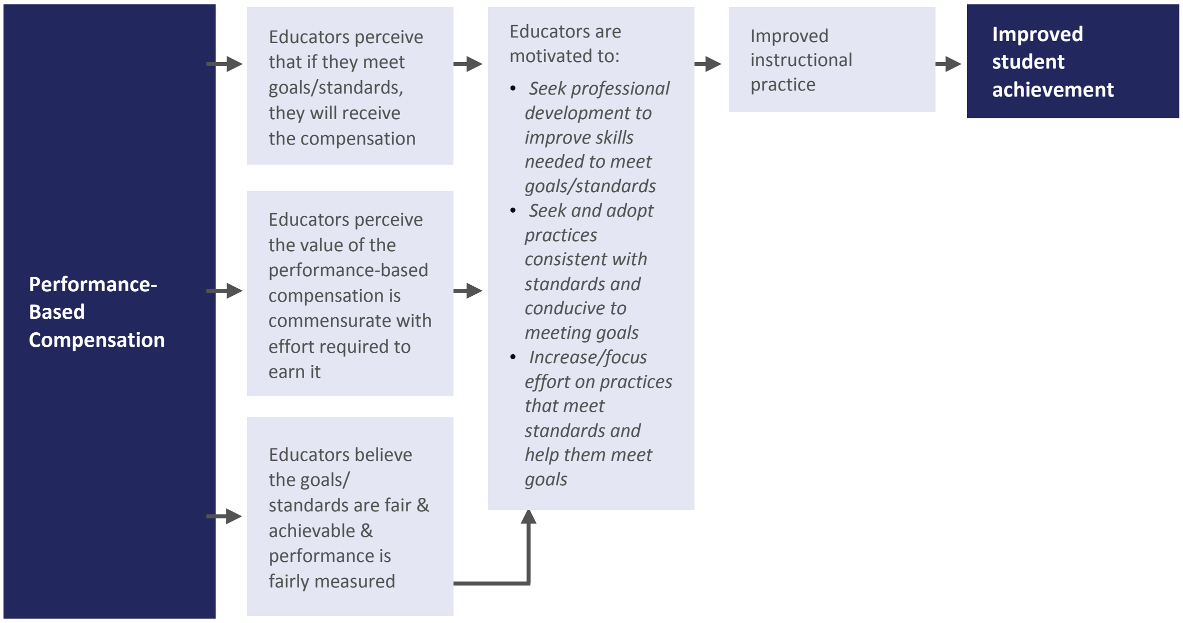
Improved student achievement

This theory of change shows the two paths along which researchers have postulated that performance-based compensation can influence student achievement. One is through improved retention of higher performers and increased attrition of low performers. The other is through motivating current educators who are less than highly effective to improve practice.

Along the first path, if the grantee appropriately designed and communicated the PBCS, then educators who are higher performers should be paid more, and lower performers less, than under the traditional compensation system. This, in turn, leads to greater retention of higher performers and lower retention of lower performers, which then leads to a higher proportion of effective and especially highly effective educators, raising the average quality of instruction. A higher average quality of instruction leads to improved student achievement.


Performance-Based Compensation
Higher pay and recognition for high performers and lower pay for low performers
• High performers stay
• Low performers leave
Higher average quality of instruction
Improved student achievement

Along the second path, if the grantee appropriately designs and communicates the PBCS, and higher performers actually receive more pay, educators will perceive that to obtain higher pay and recognition, they will need to meet the performance standards or goals, that the value of the performance-based compensation is commensurate with the effort required to earn it, that the performance goals or standards are fair and achievable, and that performance is measured fairly and accurately. These perceptions and beliefs should lead educators to seek opportunities to improve their skills, find and adapt better practices, and increase their effort or focus on practices that help them meet the standards and goals. This in turn will result in improved instructional practice and improved student achievement.

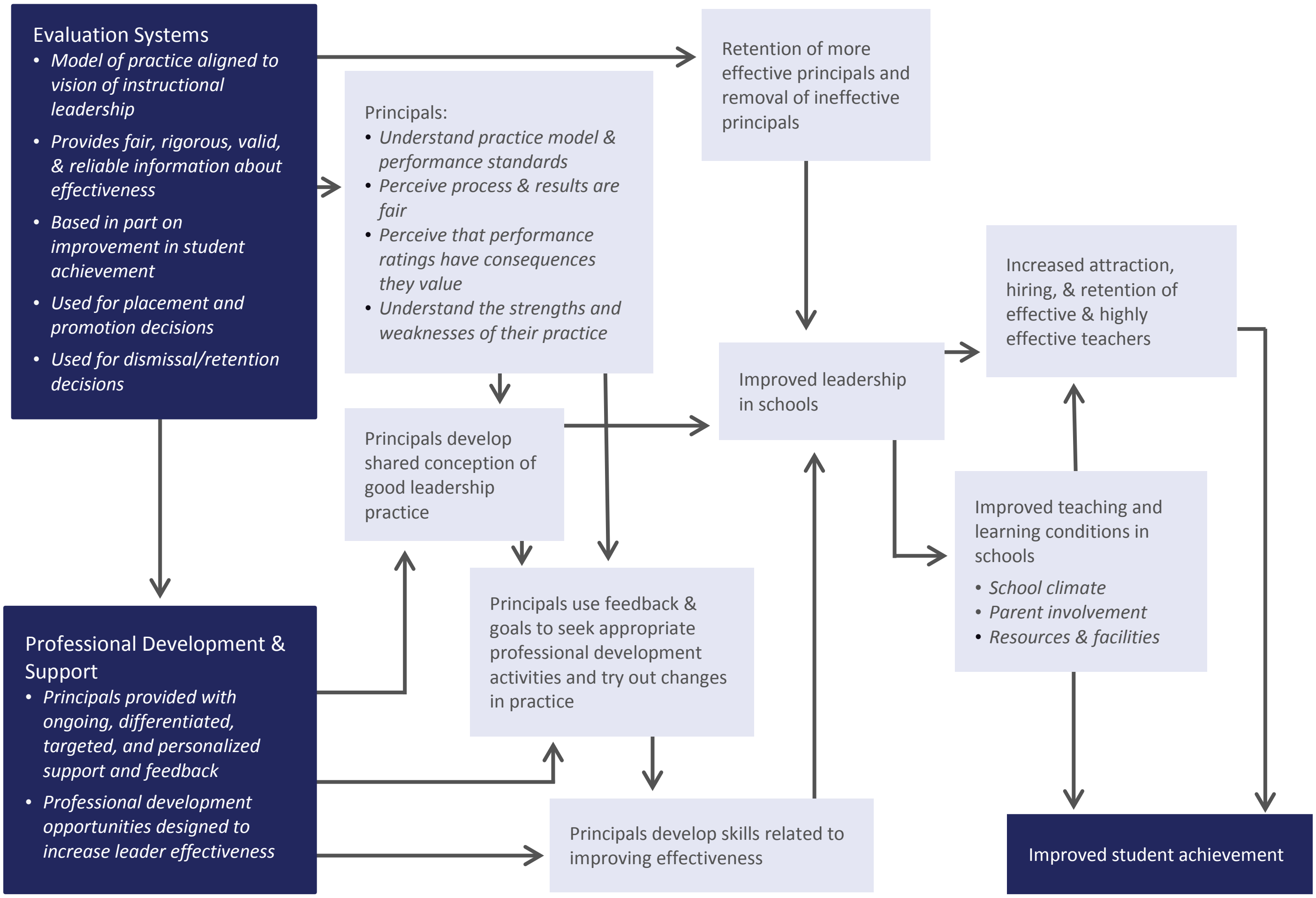

Performance-Based Compensation
Educators perceive that if they meet goals/standards, they will receive the compensation
Educators perceive the value of the performance-based compensation is commensurate with effort required to earn it
Educators believe the goals/ standards are fair & achievable & performance is fairly measured
Educators are motivated to:
• Seek professional development to improve skills needed to meet goals/standards
• Seek and adopt practices consistent with standards and conducive to meeting goals
• Increase/focus effort on practices that meet standards and help them meet goals
Improved instructional practice
Improved student achievement

Since the premise of this theory is that incentives motivate educators to modify their behavior in ways that make them more likely to receive the performance-based compensation, the diagram includes boxes that highlight requirements for motivation to occur. These include the perceptions and beliefs of educators about the contingency and value of the compensation and the fairness of the performance requirements and measurement. If grantees expect performance-based compensation to influence educator behavior, it will likely be important for evaluators to assess these perceptions. In addition, an evaluation could also examine: (a) whether educators who perceive the connection between performance and pay do in fact seek professional development, adapt new practices, and increase effort/focus on instructional practices that are likely to improve student learning; (b) whether instructional practice is in fact changing; (c) whether higher performers are retained at a greater rate that lower performers; and (d) whether student achievement is improving. This diagram could be turned into a logic model for the PBCS component by specifying program outputs such as communications to educators (e.g., percentage of educators accessing websites, number of educators attending school-level information sessions), the number or percentage of teachers receiving higher pay than they would have under the traditional pay system, and the degree of pay differentiation (e.g., differences in bonuses or salary increases between less than effective, effective, and highly effective educators).

Theory of Change for Principal Evaluation

The figure below shows a potential theory of change relating principal evaluation to improved student achievement. This theory postulates that the major pathways by which principals affect student achievement are through: (a) increased attraction, hiring, and retention of effective and highly effective teachers and (b) improving school conditions under which teachers teach and students learn, including school climate, parent involvement, and instructional resources and facilities. Improvements in school leadership affect both of these. The average quality of school leadership in a district should improve as districts retain effective principals and remove ineffective principals. Districts also improve school leadership by developing a shared conception of good leadership practice that sets the stage for a culture of high expectations for leadership, as well as improvements in the leadership skills of existing principals.

Principal evaluation leads to improved **student achievement**.


Evaluation Systems
Model of practice aligned to vision of instructional leadership
Provides fair, rigorous, valid, & reliable information about effectiveness
Based in part on improvement in student achievement
Used for placement and promotion decisions
Used for dismissal/retention decisions
Principals:
Understand practice model & performance standards
Perceive process & results are fair
Perceive that performance ratings have consequences they value
Understand the strengths and weaknesses of their practice
Retention of more effective principals and removal of ineffective principals
Increased attraction, hiring, & retention of effective & highly effective teachers
Improved leadership in schools
Principals develop shared conception of good leadership practice
Improved teaching and learning conditions in schools
School climate
Parent involvement
Resources & facilities
Principals use feedback & goals to seek appropriate professional development activities and try out changes in practice
Professional Development & Support
Principals provided with ongoing, differentiated, targeted, and personalized support and feedback
Professional development opportunities designed to increase leader effectiveness
Principals develop skills related to improving effectiveness
Improved student achievement

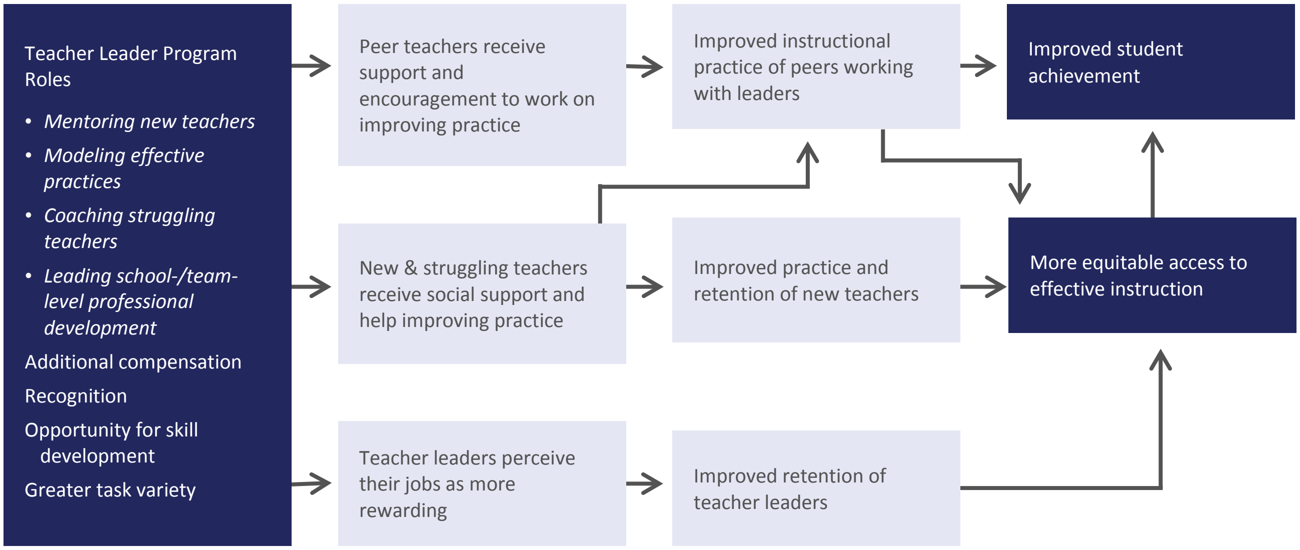
A shared conception of good practice is an intended outcome of principals' understanding of the practice model underlying the evaluation system and the performance standards and their perception of them as fair and appropriate, as well as the differentiated support and professional development opportunities they have to develop the underlying skills. If principals understand the system, perceive it as fair, perceive the ratings have consequences, and understand their strengths and weaknesses, they will be more likely to use the feedback from the evaluation process to seek appropriate professional development and try out new practices that promise to improve their effectiveness. If quality opportunities are available and principals take advantage of them, they are more likely to develop the skills needed to improve their leadership in their schools.

This theory of change could be turned into a logic model by specifying the program outputs such as communications about the evaluation process to principals, the proportion of principals evaluated and provided with feedback, the proportion taking advantage of professional development opportunities related to their evaluation feedback, and the proportion participating in coaching sessions with mentors, colleagues, or supervisors.

Theory of Change for Teacher Leadership Programs

The figure below shows a potential theory of change for relating teacher leadership programs to improved student achievement and more equitable access to effective instruction. This theory begins with defined teacher leader roles, including mentoring of new teachers, modeling effective practices to peers, coaching struggling teachers, and leading school or team professional development. (Other teacher leader roles could be defined, but we use these common roles here for illustration.) TIF-supported teacher leadership programs will typically include additional compensation and recognition, professional development relating to the roles as well as the opportunities the roles offer to develop and use new skills, and a greater variety of tasks.

Teacher leadership leads to improved **student achievement and equitable access to effective instruction**.


Teacher Leader Program Roles
• Mentoring new teachers
• Modeling effective practices
• Coaching struggling teachers
• Leading school-/team-level professional development
Additional compensation
Recognition
Opportunity for skill development
Greater task variety
Peer teachers receive support and encouragement to work on improving practice
New & struggling teachers receive social support and help improving practice
Teacher leaders perceive their jobs as more rewarding
Improved instructional practice of peers working with leaders
Improved practice and retention of new teachers
Improved retention of teacher leaders
Improved student achievement
More equitable access to effective instruction

According to this theory of change, teacher leadership programs have effects on both peers of the teacher leaders and the teacher leaders themselves. Peer teachers working with the leaders receive support and encouragement to work on improving practice through one-on-one and group professional development activities, coaching, and modeling of practice by teacher leaders.

Teacher leaders can also serve as mentors and coaches to new and/or struggling teachers, providing social support, encouragement, and specific help in improving instruction. These two pathways both lead to improved instruction, which in turn contributes to improving student achievement. New teacher mentoring can also help retain new teachers, potentially lowering turnover and reducing the tendency of high-need subjects and hard-to-staff schools from having a disproportionate number of inexperienced teachers. Teacher leadership programs may also improve the retention of teacher leaders themselves, by improving the perceived rewards of the job, ranging from more pay to greater task variety. This, in turn, can contribute to more equitable access to effective instruction, when districts retain these highly effective teachers in high-need subjects and hard-to-staff schools. More equitable access itself contributes to improved student achievement when more effective teachers teach disadvantaged students.

This theory of change could become a logic model by adding key program outputs such as the percentage of teacher leader roles filled, the number or proportion of teachers in schools that observe modeling by teacher leaders, the proportion of new teachers assigned mentors, and the number of hours of contact between mentors and new teachers. Potential teacher reactions or perceptions that short-term outcomes could include are teacher leaders' perceptions of the value of additional compensation, recognition, and development opportunities; peer or new teacher perceptions of the qualifications, credibility, and accessibility of the teacher leaders; or the usefulness of feedback, coaching, or resources teacher leaders provided.

Summary

This section has reviewed the concept of a logic model and provided an example logic model for a comprehensive TIF grant. It has illustrated how a logic model can represent key outputs and short-term outcomes that are important in assessing the degree to which grantees actually implemented intended activities, as well as medium- and long-term outcomes that they should assess in a summative evaluation. It has also provided three theories of change or theories of action that TIF grantees and their evaluators could use to develop more detailed logic models covering specific parts of their grant-supported activities. It has emphasized the importance of having a logic model or theory of change/action for evaluators to use to guide the development of evaluation questions that focus on causal links between major program elements and ultimate outcome goals, such as improved student achievement.

Resources

Borgman-Arbouleda, C. (2012). *Developing your theory of action: A facilitation guide*. New York: Action Evaluation Collaborative, www.actionevaluation.org/facil

Frechtling, J. A. (2007). *Logic modeling methods in program evaluation*. San Francisco: Jossey-Bass.

Shakman, K., & Rodriguez, S. M. (2015). *Logic models for program design, implementation, and evaluation: Workshop toolkit* (REL 2015–057). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. http://ies.ed.gov/ncee/edlabs

Organizational Research Services. (2004). *Theory of change: A practical tool for action, results and learning*. Prepared for the Annie E. Casey Foundation, www.aecf.org

W. K. Kellogg Foundation. (2004). *Using logic models to bring together planning, evaluation, and action: Logic model development guide.* https://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide

2 Developing Evaluation Questions

Once program administrators and evaluators have agreed on a logic model or set of logic models, they can use them to guide the development of the specific questions the evaluation will attempt to answer. This section provides examples that illustrate how grantees and evaluators can develop formative and summative evaluation questions using the inputs, activities, outputs, and outcomes represented in their logic models.

Collaborative Development of Evaluation Questions

The most useful evaluations incorporate questions that the program staff and stakeholders find important. Thus, program administrators and evaluators should develop evaluation questions for formative and summative evaluations collaboratively. It is worth taking the time at the beginning of an evaluation to have program administrators and representatives of key stakeholder groups both review the logic model and brainstorm questions. Evaluators can then propose a final set of questions based on resources available, Department of Education requirements (e.g., Government Performance and Results Act measures), and expected program implementation issues. Agreement on a working logic model and on evaluation questions derived from it also helps all 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 among evaluators, program staff, and stakeholders on final questions sets the stage for a successful project.

To be of maximal use, TIF evaluations should include questions that grantees can use for both formative (program improvement) and summative (assessment of program impact) purposes. Formative evaluation questions focus on whether grantees implemented TIF activities with fidelity, whether they produced expected outputs, and whether they observed intended and short-/medium-range outcomes. Program managers are often most interested in formative evaluation, using answers to evaluation questions as ongoing feedback about program implementation and to identify potential areas of improvement. For formative uses, questions about activities, outputs, and short-term outcomes are generally of most interest. Summative uses rely on information about medium- and long-term outcomes. However, summative evaluation can also benefit from attention to activities, outputs, and short-term outcomes in order to assess implementation fidelity and the degree to which grantees carried out program activities as designed. Whether they implemented a program with fidelity is increasingly seen as crucial to judgments of whether grantees can attribute any changes in outcomes to the program (O’Donnell, 2008).

In the next sections, example evaluation questions related to context, fidelity of implementation (activities and outcomes), and short-, medium-, and long-term outcomes are provided, based on the logic model and theories of change presented in Section 1.

Example Contextual Evaluation Questions

Context Factor	Possible Evaluation Questions About Program Context	Potential Impacts on TIF Components or Outcomes
State Policy	• How does the state's school accountability system align with TIF performance measures? • What are the state's policies or procedures involving student testing (e.g., content and timing of tests, which grades are tested)? Have they changed or are likely to change? • What state regulations or policies affect classroom observations (for example, number of observations, rubrics allowed, training for evaluators)? • What state regulations or policies affect educator compensation (for example, a state minimum salary schedule, whether bonuses are included in retirement benefit calculations)?	• Ability to implement a consistent measure of educator impact on student achievement, ability to measure long-term impacts on student achievement • Quality of evaluator training, how well rubric captures vision of instruction • Ability to differentiate compensation based on performance, value of bonuses to educators
District Policies & Priorities	• How much of a priority is TIF for district leaders? • Has the district introduced any instructional initiatives (e.g., new curriculum, teacher teaming)? Do these align with the educator evaluation and support system?	• May reinforce or distract from TIF vision of instruction and implementing TIF activities • May affect outcomes expected to be influenced by TIF activities
Employee-Employer Relations	• Are there provisions of collective bargaining agreements that affect educator evaluation, compensation, assignment, promotion, and retention? How have they affected the design of TIF PBCS or evaluation systems? • Is there a relationship of trust between teachers and principal, principals and district leaders, teachers and district leaders?	• Use of evaluation results for assignment, promotion, retention; ability to differentiate compensation based on performance • Educator perceptions of fairness of evaluation system, acceptance of PBCS
Fiscal Environment	• 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?	• Educator perceptions of sustainability of PBCS and evaluation/support systems • Demand for teacher leader roles
Labor Market	• How severe are shortages of teachers in high-needs subjects? • Is there sufficient supply of potentially effective or highly effective educators in the local area? • How do the pay and benefits provided to TIF educators compare to those offered by other districts in the local labor market?	• Severe shortages limit ability to attract needed teachers and thus provide more equitable access to effective instruction • Attractiveness of performance-based compensation compared to traditional system

Developing Context Questions

Since contextual factors can have important effects on program implementation and outcomes, they should be represented in the evaluation questions. During the process of developing the logic model, evaluators may have identified some of these factors. If not, evaluators may want to begin by asking program staff about what they think could hinder or help implementation, making use of the five categories of state policy, district policy and priorities, employee-employer relations, fiscal environment, and the local educator labor market. The exhibit above (Example Contextual Evaluation Questions) provides some example evaluation questions about context that build off the logic model example in Section 1.

Some important context factors will not become apparent until program staff and evaluators get into the field after grantees begin to implement TIF activities. Evaluators will likely hear about some of them from educators only during the evaluation. Thus, the evaluation design may need to incorporate at least some interviews with educators affected by the TIF activities to ask general questions about other influences they feel might affect the implementation of TIF activities, what other initiatives the district is implementing, and what other influences they perceive are affecting their behavior.

Implementation Questions

One of the most important questions an evaluator must consider is whether projects have implemented program activities effectively. Each TIF component the logic model or theory of change diagram presents numerous activities, all or most of which need to be fully implemented for the expected outcomes to result. Assessing the extent to which projects carry out these activities as intended (fidelity of implementation) both helps administrators keep program activities on track and helps evaluators judge whether observed outcomes are likely to be due to program activities or whether a lack of impact was due to poor implementation or problems with the design of the program itself. Describing implementation and its fidelity to plan is also essential to disseminate evidence-based practice.

One framework for assessing implementation fidelity suggests measuring the content, frequency, duration, and coverage of the activities (Carroll et al., 2007). Content involves the substance of the activities: what was actually done and was it done with the quality program developers intended. Frequency (the number of times educators participate in program activities) and duration (the length of time educators participate) define the “dose” of the program each educator or school receives, which is expected to influence the degree of impact. Coverage refers to the proportion of the educators or schools that experience the activities. The program outputs identified in the logic model are thus a good starting point for developing questions about implementation. Evaluators can begin with these outputs, adding others as their knowledge of how the grantees delivered activities develops.

Example Activity Implementation Questions

TIF Program Component	Possible Activity Implementation Questions
Evaluation Systems	• What percentage of educators received training on the evaluation process? • What percentage of educators received evaluations using the TIF-supported evaluation system? • What proportion of these educators received oral or written feedback after an observation? • How many observations did each educator receive? • What proportion of educators observed participated in a post conference within a week of the observation? • What proportion of tenure and retention decisions was made considering evaluation results? • What role did evaluation results play in assignment or promotion decisions (e.g., for teacher leadership roles)?
Professional Development & Support	• What proportion of educators received feedback on their performance? • How often did educators receive feedback? • Was the evaluation feedback specific, based on the rubric, accompanied by evidence, and focused on the behavior rather than the educator? • Have educators received materials that clearly explain professional development opportunities available and their links to the evaluation system and model of practice? • What proportion of educators participated in professional development opportunities linked to the evaluation system and model of practice, or targeted using evaluation system results? • What was the frequency and duration of the professional development activities?
Performance-Based Compensation	• Have educators received access to information that clearly explains the performance goals/standards, methods of measurement, and size and timing of performance-based compensation? • Did they receive this information early enough in the school year to affect their planning? • What proportion of educators received compensation increases based on performance? • How much more compensation do effective and highly effective educators receive under the PBCS than they would have received under the traditional or prior systems? • How much less compensation do less-than-effective educators receive under the PBCS than they would have received under the traditional or prior systems? • What proportion of effective or highly effective teachers teaching in high-need subjects or hard-to-staff schools received additional compensation? • How much greater was this compensation than they would have received under the traditional or prior systems and compared to what they would have earned by not working in high-need subjects or hard-to-staff schools?
Teacher Leadership	• Have educators received information that clearly explains the leadership roles, qualifications, and hiring process, professional development and supports provided, and rewards? • What proportion of teacher leader roles was filled? • What proportion of schools had teacher leaders with various roles? • How much time do teacher leaders actually spend in leadership activities? • What proportion of eligible teachers participated in professional development opportunities that teacher leaders facilitated? • How many times did new teachers work with teacher leader mentors? • What proportion of science teachers in each school had access to a science instructional coach? • How much greater was the compensation provided to teacher leaders than they would have received without taking on a leadership role?

Short-Term Outcomes

Many of the activities of a TIF-supported human capital management system aim to motivate and support educators to change their practices to improve performance. Performance improvement is affected by educators' beliefs about the program, their abilities and resources to improve, and their professional needs and values. Thus, many of the short-term outcome questions should be about how the program outputs affect educator perceptions and beliefs and how these relate to motivation.

Example Short-Term Outcome Questions

TIF Program Component	Possible Educator Perception Questions
Evaluation Systems	• Do educators understand the practice model and performance standards, how the evaluation process works, and the consequences of evaluation ratings? • Do educators accept the practice model or performance standards as an appropriate and attainable standard of practice? • Do educators believe that the way performance is measured is fair? • Do educators understand the strengths and weaknesses of their practice and/or areas in which they could improve? • Do educators believe they have the skills, resources, and supports they need to meet the performance standards and/or improve performance? • Do educators believe that if they improve their practice or performance, their evaluation ratings will improve? • Do educators perceive that the evaluation process has increased their workload and level of stress?
Professional Development & Support	• Do educators believe that the feedback they receive is credible? • Are educators aware that professional development targeted to evaluation results is available? • Do educators perceive that the support they receive is relevant to their practice, applicable in their classrooms or schools, and will help them improve performance? • Do educators understand how they can use the feedback, professional development, and other resources to improve practice/performance? • Do educators believe that they can apply the practice changes they learn about or skills they develop in their schools or classrooms?
Performance-Based Compensation	• Do educators understand the performance goals, performance measures, and the PBCS payouts or pay increases? • Do educators accept the performance goals or standards as appropriate and attainable? • Do educators believe that the way performance is measured is fair? • Do educators perceive that the amount of performance-based compensation is commensurate with the extra work or effort required to earn it?
Teacher Leadership	• Do teachers perceive that assistance from teacher leaders is available? • Do educators perceive the teacher leaders to be credible and approachable? • Do educators perceive the assistance they receive from teacher leaders is useful and applicable to their practice? • Are potential teacher leaders aware of the leadership opportunities available? • Do potential teacher leaders perceive that the rewards of taking on the role outweigh the costs? • Do teacher leaders perceive they have sufficient resources and supports to carry out their roles (including support from other school leaders)?

As a first step on this path, evaluators can ask whether educators understand key aspects of the program components, including the performance goals or standards, how performance is measured and rewarded, and what supports are available to improve performance. Evaluators may also want to ask about whether educators perceive that the performance goals and/or practice standards are appropriate (i.e., legitimate, attainable, and consistent with other goals set by the state, district, or school). Without such acceptance, motivation to work toward TIF-related goals is not going to be strong. Perceptions about fairness of procedures and outcomes are also likely to relate to motivation. The exhibit above (Example Short-Term Outcome Questions) shows some potential evaluation questions about educator perceptions and beliefs.

Questions about short-term outcomes do not have to be limited to attitudes or perceptions. For example, with regard to professional development and support, evaluators can also ask whether professional development records show an increased demand for courses related to practice change, whether teachers have requested help from coaches or mentors on skills related to the evaluation rubric, or how many educators' professional development plans include references to improving areas of need identified by evaluation results. With respect to teacher leadership programs, it could be useful to find out whether teachers working with leaders try out the new practices or use resources suggested by coaches or mentors, how many candidates applied for each teacher leader opening, or how teacher leaders participate in school leadership teams.

Medium-Term Outcomes

The logic model in Section 1 illustrated several medium-term outcomes that could follow from the short-term outcomes and lead to improved student outcomes, including to the development of a shared conception of good practice, improved instructional and leadership practices, increased retention of effective and highly effective teachers and school leaders, and lower retention of those who are less than effective.

Developing evaluation questions about a shared conception of practice begins with understanding the vision of instruction underlying the teacher and administrator evaluation systems. It can be useful to identify several key propositions about instruction and leadership that represent the most important and distinctive aspects of the vision. Reviewing the professional practice rubrics can help identify these propositions.

Potential evaluation questions about the development of a shared conception of good practice could include:

- To what extent do teachers, teacher leaders, and administrators agree that the key components of the vision of instruction underlying the evaluation systems represent the instruction teachers should be trying to implement and leaders to support?
- Has the level of agreement changed over the course of the TIF project?
- Do teachers, teacher leaders, and administrators reference these aspects in their conversations about instruction?
- Do district or school communications about instruction other than those related to the evaluation system reference these aspects?

- To what extent do educators perceive that their colleagues share the same assumptions and values about instruction or leadership?
- Has this perception increased over the course of the TIF project?

The second medium-term outcome in the example logic model is improved professional practice. Change in instructional practice is a key outcome to assess since the evaluation and support systems central to TIF human capital management systems are based on a vision of instruction that will improve student achievement. Leadership practice, in turn, is a key support for improving instruction (Hitt & Tucker, 2016; Leithwood & Louis, 2012). Ideally, increasing the degree to which instruction (and the leadership practice that supports it) realizes this vision is a major goal of the TIF project. However, assessing changes in practice can be extremely complex. While an in-depth discussion of assessing changes in educator practice is beyond the scope of this guide, evaluators should work with program administrators to understand the vision of instruction and leadership the TIF project intends to support, including the practices it expects teacher leaders to model or support. In this case, the observation rubric or rating scale may provide a good place to start in developing specific questions about instructional practice change. Grantees intend TIF evaluation and support systems to support change in practice toward the model underlying the rubrics. The exhibit below shows examples of questions about instructional and leadership practice.

Example Medium-Term Outcome Questions

Practice	Possible Evaluation Questions
Instructional Practice	• Have teachers and schools aligned the curriculum to the tests used to measure student achievement growth or to underlying state standards? • To what extent does instruction focus on state content standards? Has this degree of focus increased during the TIF project? • To what extent has the use of student assessment results and other student information for instructional planning and differentiation increased? • To what extent have teachers increased the use of instructional techniques that encourage student engagement? • To what extent has teachers' use of techniques to improve classroom climate and develop positive relationships with students increased?
Leadership Practice	• Have school leaders increased their emphasis on recruiting and selecting effective and highly effective teachers? • Have school leaders adapted school professional development plans to the strengths and weaknesses of their faculties, as shown by teacher evaluation results? • Have school leaders increased their communication with staff about the school mission and strategies for improving student achievement? • Have leaders implemented practices to improve school climate/culture? • Are school leaders using data to make decisions, and has this use increased during the TIF project?

It is important to remember that state or district policies outside of TIF will influence some changes in practice and that evaluators cannot completely anticipate these changes when designing questions. Thus, they should also include general questions about any other practice changes teachers or school leaders have made since the initiation of the TIF program.

One way to assess practice change is to track whether the performance ratings of teachers and school leaders made using the evaluation rubrics has improved over time, both on average and longitudinally for specific educators who had a less than highly effective rating at the beginning of the project. However, it is important to recognize that changes in ratings may not always accurately represent changes in practice. Because these ratings have consequences for educators, raters may have motives for leniency and to show performance has improved. Further, reports about the distribution of ratings from prior rounds of TIF and from other sources suggest that relatively few educators receive a less-than-effective rating, making it hard to see change. It could be advisable to develop and use additional, independent measures of practice, unconnected to the consequences of the human capital management system, for at least a sample of teachers or principals. Measures could include co-observations or ratings of videos by independent raters, instructional practice logs, classroom walk-throughs, or reviews of artifacts like student work.

These independent assessments could focus on the most important aspects of practice and/or the aspects the evaluation and support activities are most likely to influence.

The third medium-term outcome shown in the logic model in Section 1 (and the theory of change for principal evaluation) is improved teaching and learning conditions. The logic model shows these as medium-term outcomes because they affect other medium-term outcomes such as teacher retention and improved instructional practice of teachers, as well as support long-term improvements in student achievement. Teaching conditions refer to the context of teaching as teachers experience it, and there are many different conceptions of what the most important dimensions of the teaching context are (Ladd, 2011; Protheroe, 2011; Berry et al., 2008). When developing specific evaluation questions about teacher working conditions, it is advisable to work with program administrators and stakeholder groups to identify what aspects of working conditions TIF activities are expected to influence (such as school leader evaluation and professional development) and agree on which aspects are most likely to affect teacher retention and openness to improving instruction.

Similarly, learning conditions refer to a variety of school and classroom policies, practices, and characteristics students experience, such as safety, discipline, student assignment and grouping, and access to technology (Beresford, 2003; Kutsyuruba, Klinger, & Hussain, 2015). Again, it is advisable to work with program administrators and stakeholder groups to identify what aspects of student learning conditions grantees expect TIF activities to influence and to affect student achievement.

The fourth medium-term outcome the logic model shows is the increased retention of effective educators, with particular focus on those in hard-to-staff schools and high-need subjects. Evaluation questions related to this outcome could include:

- Have the retention rates of effective and highly effective educators increased over the period of the TIF project?

- What is the retention rate of less-than-effective educators, and has it decreased over the period of the TIF project?
- What are the retention rates of effective and highly effective educators in high-need subjects and hard-to-staff schools?
- Have these retention rates increased over the period of the TIF project and in comparison to rates for similar educators in other subjects and schools?

Evaluators should be able to use the district human resources information system data to find out whether those evaluated as better performers were more likely to stay, controlling for age, experience, and other factors known to influence teacher turnover. It may also be fruitful to ask teachers and administrators if they perceive that 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.

Long-Term Outcomes

The two long-term outcomes the example TIF logic model depicted were improved student achievement and more equitable access to effective instruction. These are the overall goals of the TIF grant program, and evaluation designs should assess them, both to ensure that the Department and federal policymakers know about the success of TIF and to inform considerations of whether states should sustain TIF-supported activities after the grant expires.

Considering first improved student achievement, there are several ways to pose evaluation questions. First, evaluators can ask whether test scores, percentages of students scoring proficient or above on state tests, and other student outcomes such as graduation rates and college/career entrance have improved. While there are important issues to consider when holding schools and educators accountable for educational attainment indicators like test results, improving them is one of TIF's overall goals. Showing that these "bottom line" outcomes have improved can also be important in convincing stakeholders to sustain TIF-supported activities after the grant period.

A more nuanced question is whether the net value-added of the education system has improved. If TIF-supported activities to improve the quality of the educator workforce have been successful, we would expect to see that the productivity of the system has improved. Are students achieving more given their starting points and situations (e.g., poverty, English learner status)?

Assessing whether the average value-added has increased at the student level could address this question. The aim here is not to attribute student achievement to a particular teacher or school, but simply to compare student-level value-added before and after the implementation of TIF-supported activities. A recent evaluation of the Denver ProComp initiative provides an example of how this might be done (Goldhaber & Walch, 2012).

As is discussed later in this guide (Section 4, Evaluation Design 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 can and should at least track trends in outcomes and compare the trend after TIF implementation to the trend before implementation (Milanowski & Finster, 2016). District stakeholders will typically want to know if student outcomes are improving. Evaluators can also make some judgments about program impact even without a strong evaluation design if the evaluation has addressed all the elements of the logic model (Trochim, 1985, 1989). For example, if the evaluation has found (a) faithful implementation, (b) few contextual influences, (c) expected educator reactions to the activities, and (d) changed instructional practice, as well as improved student outcomes, then decisionmakers can be relatively comfortable attributing some of an observed upward trend in student achievement to TIF grant activities. On the other hand, if the evaluation has found faithful implementation, but no changes in achievement, this would be evidence that the program had minimal impact.

Moving on to more equitable access to effective instruction, one precursor of more equitable access is an increase in the proportion of effective and highly effective educators working in hard-to-staff schools and high-need subjects. Thus, whether these proportions have increased in these schools and subjects is a likely evaluation question. Another approach would be to track changes in the proportions of disadvantaged students served by effective educators. More nuanced questions about equitable access could include:

- Has the probability of a less-than-effective teacher teaching a disadvantaged student for more than 1 year declined? Has the probability of such students receiving instruction by a highly effective teacher for more than 1 year increased?
- Has the effect of student demographic characteristics like race and poverty on student achievement found in typical value-added models decreased over the course of the TIF grant?

It could be useful to look more specifically within subjects (e.g., comparing trends in proportions of effective/highly effective math, science, and reading/English language arts teachers) in schools serving disadvantaged versus advantaged students. Variation across subjects could indicate that TIF activities might have to be more differentiated or targeted. Evaluators may also want to review state equity plans to identify questions suggested by state priorities and other measures of equity that might be worth tracking over time. Evaluators might also want to consider asking about whether achievement gaps themselves are narrowing, using state or district assessment results.

Unintended Consequences

In addition to input, activity, output, and outcome questions, it is important for evaluators to consider the possibility of unintended or unexpected consequences of TIF program activities. Articles by Morrell (2005) and Manley (2013) discuss methods to identify potential unintended consequences. Research on accountability and performance incentives suggests that there are common types of unintended consequences such as gaming the system and increasing turnover of those who do not receive awards. Evaluators thus 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?
- Has cooperation and collegiality among teachers decreased?

- Have highly effective principals become more unwilling to move to low-performing schools due to fear of lower evaluations or compensation?
- Has turnover of effective teachers or principals increased due to perceived additional burdens of more rigorous evaluation?

Finding unintended consequences does not necessarily mean that grantees should discontinue TIF activities or that they are the wrong path to improving student achievement. Some unintended consequences may be due to shortcomings in implementation, as well as features of program design or conflicts with other state or district policies or initiatives. Surfacing potential unintended consequences early in the implementation process allows for program staff to redesign activities and improve problematic implementation. Open-ended interview and focus group questions about how the activities are being perceived can provide a valuable early warning of potential unintended consequences and are therefore worth including in the evaluation design.

Program Evaluation and Sustainability

Program evaluation can make an important contribution to sustaining TIF activities during and after the grant period. TIF technical assistance providers have developed a four-component approach to promoting sustainability (Pasley, Keheler, & Gould, 2015):

- Increasing stakeholder support and communication;
- Building capacity for quality programs;
- Developing financial support and ongoing funding for efforts;
- Understanding and communicating return on investment, including how instructional practice and student achievement have improved.

Evaluations can both assess the quality of communication and shareholder support as well as contribute to it by providing accessible, unbiased information about how grantees are implementing TIF activities and whether the activities are having the expected outcomes. By providing formative information on program implementation and short- to medium-term impacts, the evaluation helps program managers monitor program quality. By providing information on the impact of TIF activities on long-term outcomes, the evaluation can help program managers and stakeholders understand whether practice and achievement have improved and provide a key input to assessing the return on the investment in TIF activities.

Evaluators can develop evaluation questions to specifically address these sustainability components. For example, consider communication and stakeholder engagement. TIF technical assistance providers have recommended that grantees develop a formal communications plan. How well they implement this plan and their success in communicating it could be addressed based on the questions in the exhibit below.

Example Activity, Output, and Outcome Questions

Questions About Activities and Outputs	Questions About Outcomes
• Was a communication plan developed to explain key TIF components to teachers, principals, and other stakeholders (e.g., the vision of instruction, performance standards/goals, evaluation procedures, and performance-based compensation)? • Were the activities the plan detailed carried out?	• How well do teachers, principals, and other stakeholders understand the TIF components? • What components of TIF do stakeholders value most or wish to continue?
• Were all of the district departments (e.g., payroll, professional development, human resources, information technology) involved in each TIF component provided with the information they needed about their roles and how they were expected to support TIF?	• How well do staff in these departments understand their roles in TIF? • Did they carry out these roles as planned or expected?
• Did stakeholders receive regular updates on the progress of implementation?	• How well do teachers, principals, and other stakeholders understand the status of the TIF program?

Educator perceptions about the fairness of performance measurement, the value of performance compensation, and about workload, autonomy, stress, and the value of interactions with teacher leaders are also likely to affect their support for continuing TIF components, so many of the evaluation questions, as shown in the example above, are also relevant to sustainability.

Understanding and communicating return on investment requires not only an assessment of the medium- and long-term impacts of TIF activities, but also an assessment of their costs. Sustaining TIF activities also requires estimating likely costs and identifying potential funding sources to keep them going after the grant period. Evaluators might therefore also want to include questions about the costs of TIF activities in their evaluation plans. Once costs have been established, grantees can determine return on investment and cost effectiveness, as discussed further in Section 3 of this guide.

> ***"Understanding and communicating return on investment requires not only an assessment of the medium- and long-term impacts of TIF activities, but also an assessment of their costs."***

Resources

Beresford, J. (2003). Developing students as effective learners: The student conditions for school improvement. *School Effectiveness and School Improvement, 14*(2), 121-158.

Berry, B., Smylie, M., & Fuller, E. (2008). *Understanding teacher working conditions: A review and look into the future.* Carrboro, NC: Center for Teaching Quality. http://www.teachingquality.org/sites/default/files/UnderstandingTeacherWorkingConditions-AReviewandLooktotheFuture.pdf

Carroll, C., Patterson, M., Wood, A., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, 40.

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review, 31*(6), 1067-1083.

Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: A unified framework. *Review of Educational Research, 86*(2), 531-569.

Johnson, S. M. (2006). *The workplace matters: Teacher quality, retention, and effectiveness*. Working Paper. Washington, DC: National Education Association Research Department.

Kutsyuruba, B., Klinger, D. A., & Hussain, A. (2015). Relationships among school climate, school safety, and student achievement and well-being: A review of the literature. *Review of Education, 3*(2), 103-135.

Ladd, H. F. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis, 33*(2), 235-261.

Leithwood, K., & McAdie, P. (2007). Teacher working conditions that matter. *Education Canada, 47*(2), 42-45.

Leithwood, K., & Louis, K. S. (2012). *Linking leadership to student learning.* Jossey-Bass, San Francisco, CA.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Manley, R. A. (2013). The policy delphi: A method for identifying intended and unintended consequences of educational policy. *Policy Futures in Education, 11*(6), 755-768.

Milanowski, A., & Finster, M. (2016). *Ways to evaluate the success of your teacher incentive fund project in meeting TIF goals.* Rockville, MD: Westat. Prepared for the U.S. Department of Education Teacher Incentive Fund.

Morell, J. A. (2005). Why are there unintended consequences of program action, and what are the implications for doing evaluation? *American Journal of Evaluation, 26*(4), 444-463.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Pasley, J., Keleher, J., & Gould, T. (2015). *Sustaining your TIF efforts: A reflection guide.* https://www.tifcommunity.org/sites/default/files/resources/tif_paper_sustaining_draft4.pdf

Protheroe, N. (2011). *Workplace conditions that matter to teachers. principal's research review: Supporting the principal's data-informed decisions,* vol. 6, no. 1. Reston, VA: National Association of Secondary School Principals.

Trochim, W. M. K. (1989). Outcome pattern matching and program theory. *Evaluation and Program Planning, 12*(4), 355-366.

Trochim, W. M. (1985). Pattern matching, validity, and conceptualization in program evaluation, E*valuation Review, 9*(5), 575-604.

3 Using Qualitative, Quantitative, and Mixed Methods

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. For example, evaluators can use qualitative methods such as interviews or focus groups to gain insight into how participants understand program components, allowing design of effective survey questions that can be administered to a larger and more representative group. Evaluators can also use interviews and focus groups to interpret unexpected patterns in survey responses and identify unintended consequences.

Using a Qualitative Methods Approach

Qualitative methods are best suited for in-depth exploration of relationships between program components and participants' reactions and behaviors. 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 ensure that the logic model captures all program components, outputs, and contextual factors.

TIF evaluations can collect qualitative evidence from a variety of sources. While evaluators often use interviews and focus groups, it may also be useful to review documents, including:

- Proposals;
- Budgets, plans, and program descriptions;
- Newspaper articles;
- Memos and written communications
- Meeting agendas and minutes.

Once a district collects the necessary qualitative data, the next step is data analysis. 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 measures of the constructs represented in the logic model and assess relationships among them using statistical analysis. For example, an analysis could assess the relationship of teachers' perceptions of the fairness of the evaluation system and the credibility of their evaluators to their reported use of evaluation feedback in changing practice. Ultimately, some form of quantitative analysis of changes in outcomes such as student achievement or teacher retention will also be needed, as discussed in the next section.

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).

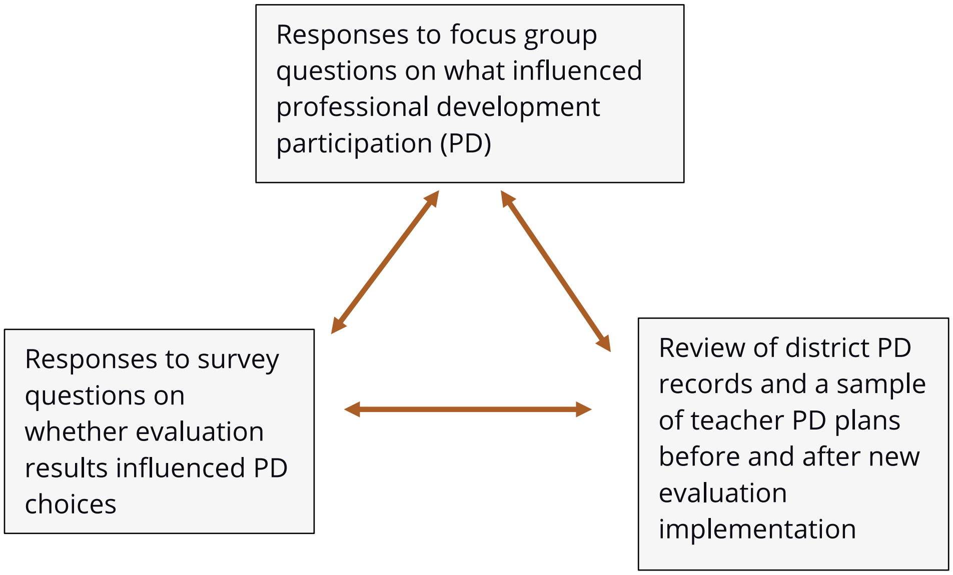
A mixed-methods approach to evaluation uses the strengths of both quantitative and qualitative methods to achieve systematic, comprehensive, and dependable findings (National Science Foundation, 1997). It is important that the designers of a mixed-methods approach select the appropriate combination of 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).

For example, we would be more confident concluding that the introduction of a new evaluation system influenced teachers’ professional development choices if each of the multiple ways we explored that question gave the same result.

Mixed Methods for Triangulating Conclusions

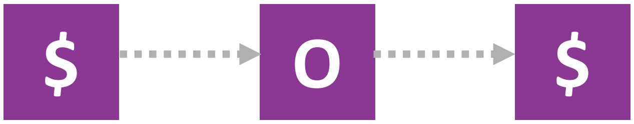
Has the evaluation system encouraged teachers to seek professional development related to improving practice?


Responses to focus group questions on what influenced professional development participation (PD)
Responses to survey questions on whether evaluation results influenced PD choices
Review of district PD records and a sample of teacher PD plans before and after new evaluation implementation

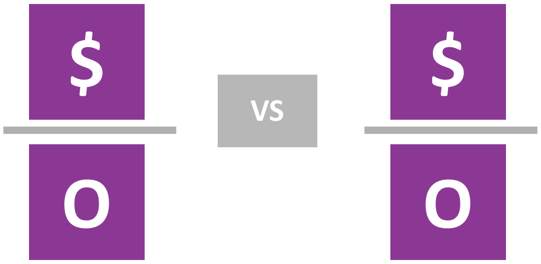
A crucial activity of mixed-methods research is synthesizing the information from quantitative and qualitative analysis. Using the triangulation approach, evaluators use multiple information sources to support the validity of their conclusions and ultimately increase policymakers' confidence in using results for decisionmaking (Shufflebeam & 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

Cost-benefit



Cost-effectiveness



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 a state education agency (SEA), local education agency (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:

- Is the program producing benefits sufficient to justify the costs?
- 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 & 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 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 & McEwan, 2001).

Resources

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.

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, & K. E. Newcomber (Eds.), *Handbook of practical program evaluation* (pp. 456-488). San Francisco: Jossey-Bass.

Levin, H. M., & McEwan, P. J. (2001). Cost-*effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

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 & L. Sharp (Eds.). Washington, DC: Author.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, & applications*. San Francisco: Jossey-Bass.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage Publications, Inc.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures* (Revised ed.) Thousand Oaks, CA: Sage Publications, Inc.

4 Evaluation Design Selection Framework

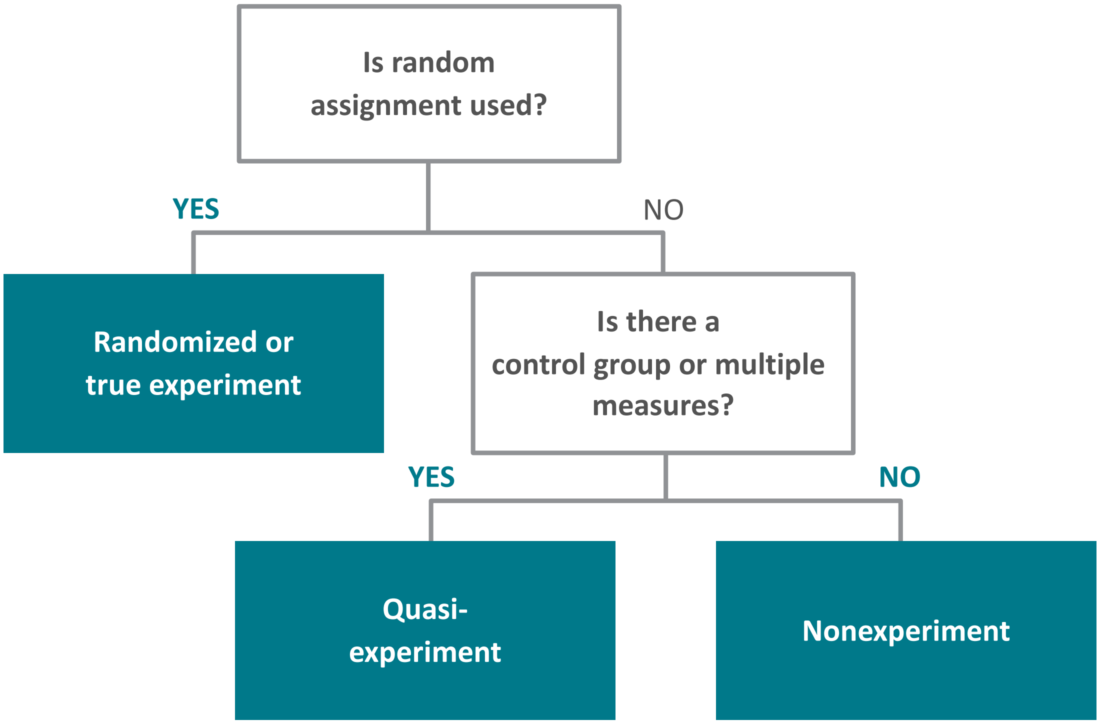
This section focuses on choosing the appropriate design for the evaluation of TIF impacts, considering experimental, quasi-experimental, and nonexperimental designs. More specifically, the section discusses the requirements for each design and the strength of each at establishing causal relationships between an intervention and an outcome.

Moreover, the section addresses the importance of evaluators considering the type of program and available data when selecting the evaluation design 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, & Wright, 1979; Teddlie & 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 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.)

Random assignment, a control group, and multiple measures help guide your **design decisions.**


Is random assignment used?
YES
NO
Randomized or true experiment
Is there a control group or multiple measures?
YES
NO
Quasi-experiment
Nonexperiment

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 (RCT) experiments, quasi-experiments, surveys using representative samples, and cohort/cross-sectional samples (Rossi et al.,1979; Teddlie & Tashakkori, 2008). The goal of achieving the greatest degree of internal and external validity should guide the selection from these approaches. This section focuses primarily on experimental and quasi-experimental designs, but a number of resources are available that discuss additional designs (Campbell & Stanley, 1963).

Treatment and Control Evaluation Designs

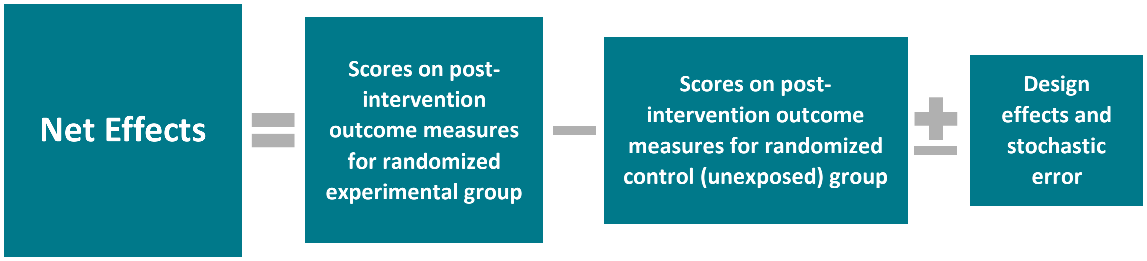
Randomized controlled experimental evaluation designs provide the strongest internal and external validity and, consequently, the most credible information about program outcomes. Not surprisingly, policymakers, stakeholders, and the general public often prefer these types of designs because they tend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

RCTs 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

This is the logic equation for **randomized treatment and control**:



Net Effects

=

Scores on post-intervention outcome measures for randomized experimental group

−

Scores on post-intervention outcome measures for randomized control (unexposed) group

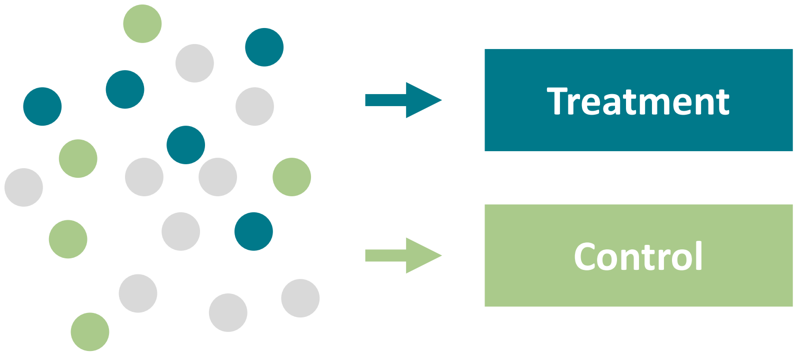
±

Design effects and stochastic error

Many consider the 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 & 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 & Springer, 2007). For an example of a TIF evaluation using an RTC design (Glazerman & Seifullah, 2012).

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. 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 for participation and control.

Several programs to accomplish this are free, including PowerUp! (Dong & Maynard, 2013), Gpower (Buchner, Erdfelder, & Faul, 1996), and Optimal Design (Spybrook, Raudenbush, Congdon, & 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 & 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.

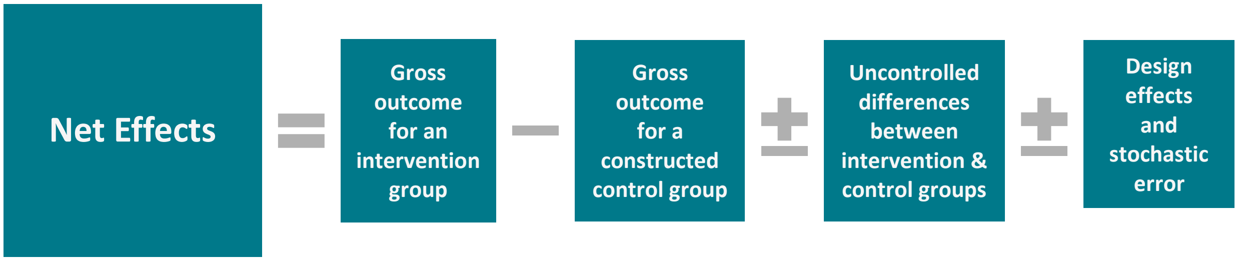
Partial-coverage programs: Randomized experimental designs are applicable only to partial-coverage programs in which there are sufficient numbers of nonparticipants from which to draw a control or comparison group.

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

This is the logic equation for **quasi-experimental design:**



Net Effects = Gross outcome for an intervention group − Gross outcome for a constructed control group ± Uncontrolled differences between intervention & control groups ± Design effects and stochastic error

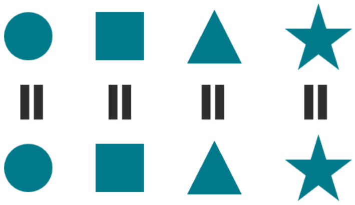
Quasi-experimental designs are quantitative outcome designs that do not involve randomly assigned comparison groups. Evaluators usually select this type of evaluation because either the assignment to intervention and control condition is not within the evaluator's control or because of political, ethical, 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 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 & Glass, 1968).

Constructing Control and Comparison Groups in Quasi-Experimental Evaluation

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 address this, at least partially. 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 RCT 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.

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 evaluators can describe the relationships among student and school characteristics, program participation, and outcomes 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.

Nonexperimental Designs

Since nonexperimental 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 nonexperimental 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 nonexperimental 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.

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.

Resources

Boruch, R. F. (2005). *Place randomized trials: Experimental tests of public policy*. Thousand Oaks, CA: Sage Publications, Inc.

Bracht, G. H., & Glass, G.V. (1968). The external validity of experiments. *American Educational Research Journal, 5*(4), 437-474.

Buchner, A., Erdfelder, E., & Faul, F., (1996). G Power: A general power analysis program. *Behavioral Research Methods, 28*(1), 1-11.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin Company.

Cook, T. D., & Campbell, D. T. (1979). *Quasi- experimentation: Design and analysis issues for field settings*. Florence, KY: Cengage Learning.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness, 6*(1), 24-67.

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years*. *Final Report*. Washington, DC: Mathematica Policy Research Inc.

Nave, B., Miech, E. J., & Mosteller, F. (2000). A rare design: The role of field trials in evaluating school practices. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan, (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.). Boston: Kluwer Academic Press.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909-949.

Rossi, P. H., Freeman, H. E., & Wright, S. R. (1979). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage Publications, Inc.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009*). Optimal design for longitudinal and multilevel research: Documentation for the optimal design*. Software V.2.0.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research.* Thousand Oaks, CA: Sage Publications, Inc.

Teddlie, C. B., & 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.

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 report their evaluation findings effectively. 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 decision makers 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 & Shinkfield, 2007).

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 & 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 clear, relevant, forceful, and convincing to stakeholders. The Joint Committee on Standards for Educational Evaluation's Program Evaluation Standards (Yarbrough, Shulha, Hopson, & Caruthers, 2011) emphasizes the importance of relevance to a variety of stakeholders by being comprehensive, clear, timely, and balanced (see Appendix 2). 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 on 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.

Resources

Gangopadhyay, P. (2002). *Making evaluation meaningful to all education stakeholders*. Retrieved December 2010, from http://www.wmich.edu/evalctr/archive_checklists/makingevalmeaningful.pdf

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, & applications*. San Francisco: Jossey-Bass.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

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 an 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.

It is important to know the **signs of resistance** to evaluation tactics.

- **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.

Both **individual** and **contextual** factors work together to cause **evaluation anxiety**.

INDIVIDUAL SOURCES

- Lack of experience with program evaluation;
- Negative past experiences with program evaluation;
- Excessive ego involvement with program model;
- Excessive fear of negative consequences.

CONTEXTUAL SOURCES

- Failure to highlight program accomplishments;
- Social norms;
- Role ambiguity.

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. The source above summarizes the specific causes of evaluation anxiety at the individual and contextual levels, also referred to as the evaluation anxiety construct.

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. 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.

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:

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

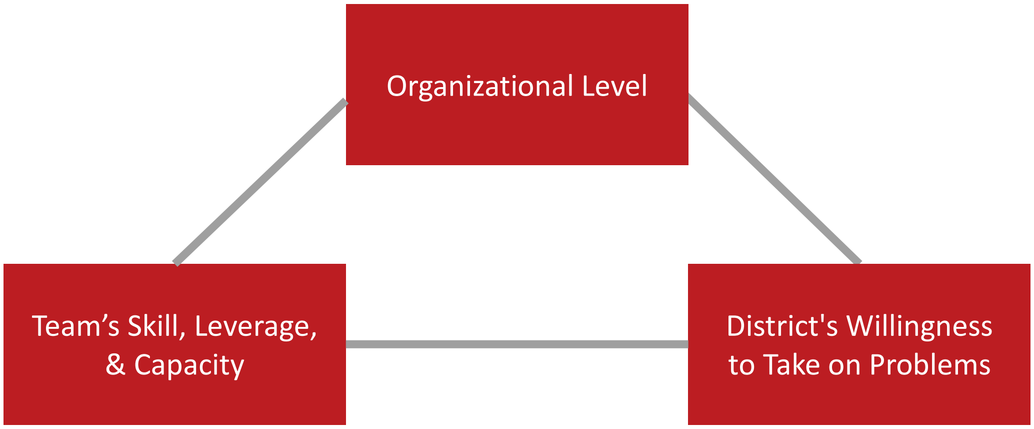
This question speaks to the appropriateness of doing the formative or the summative evaluation internally. Stufflebeam (2002) lists the following expertise as necessary for an internal evaluation unit: field work, group process, interviewing, measurement, statistics, surveys, cost analysis, values analysis, 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.

> ***"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."***

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.

When planning an evaluation internally, consider the following things:


Organizational Level
Team's Skill, Leverage, & Capacity
District's Willingness to Take on Problems

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.

How do you address **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.

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 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:
  - Did TIF improve student achievement by increasing teacher and principal effectiveness?
  - How well did stakeholders understand the new compensation systems?
  - How much "buy-in" did the TIF program have from the various stakeholders?
  - How did TIF change the allocation of effective teachers across schools?

- Evaluators conducting a formative evaluation might also need to answer myriad additional questions, such as:
  - 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?
  - How congruent is the espoused program logic model with the actual program in action?

The RFP should include 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. Such contracts at a minimum should define:

- The evaluator's audience;
- The evaluation questions;
- The substance of interim and final reports;
- 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 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 step of the evaluation design. Stufflebeam has developed a useful checklist for constructing an evaluation budget (Stufflebeam & Shinkfield, 2007).

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.

Resources

Donaldson, S. I., Gooler, L. E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation, 23,* 261-273.

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.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Stufflebeam, G., Madaus, F., & Kellaghan, T. (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.). Boston: Kluwer Academic Press.

Stufflebeam, D. (1999). *Evaluation contracts checklist*. Accessed September 9, 2010, from www.wmich.edu/evalctr/checklists/

Stufflebeam, D. (2001). The meta evaluation imperative. *American Journal of Evaluation, 22,* 183-209.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, & applications.* San Francisco: Jossey-Bass.

Volkov, B., & King, J. (2002). *A checklist for building organizational capacity*. Accessed October 20, 2010, from http://www.wmich.edu/evalctr/archive_checklists/ecb.pdf

APPENDIX 1
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 another. 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.

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.

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.” 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.

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 threat to external validity is 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
Program Evaluation Standards

The Joint Committee on Standards for Educational Evaluation has developed a set of program evaluation standards that both evaluators and grantees can use in planning evaluations and reviewing the quality of evaluation reports. The work of the joint committee, and the standards themselves, are described at: http://www.jcsee.org/. The standards cover almost all aspects of evaluation. The most current version of the standards is: Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

When choosing an evaluator or considering a proposed evaluation design, grant managers can use the Standards to think about specifications for the project and questions to ask the evaluators, to promote getting the best quality evaluation possible. Among the topics covered by the standards are:

- Utility—the extent to which program stakeholders find evaluation processes and products valuable in meeting their needs;
- Feasibility—covering how the evaluation project is designed and managed;
- Propriety—covering fairness, justice, and legality, conflicts of interest, and reporting of findings;
- Accuracy—the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support judgments about program impacts; and
- Accountability—covering the adequacy of documentation of evaluations and perspective focused on improvement and accountability for evaluation processes and products.