**“How Do We Know” … if Our Program is Working?**

**An Introduction to Program Design and Program Evaluation**

**Introduction**

The “[How Do We Know?](http://www.doe.mass.edu/research/howdoweknow/)” initiative supports The Massachusetts Department of Elementary and Secondary Education (ESE), districts, and schools to build, use, and share evidence of impact and efficacy. A program evaluation is a tool that will help you build evidence about whether your program’s efforts are implemented as intended and having the desired impact. Evaluation can also help identify areas for improvement. This guide provides you with information and resources about program planning and evaluation that will help prepare you to embark more knowledgeably on the design of both programs and evaluations. The purpose of this guide is NOT to turn you into an evaluation expert or to make program staff into researchers. Rather, the guide provides an overview of the major steps and considerations in program and evaluation design, including using logic models, to design programs and evaluations, along with additional resources for learning more. This guide will be useful to ESE, district and school program staff members who are new to evaluation, who have been asked to manage or commission a program evaluation, or who provide support to districts or schools that are designing programs and evaluations.

ESE, districts, and schools sometimes miss opportunities to measure implementation and impact because we don’t plan for evaluation early enough. The ESE Office of Planning and Research (OPR) hopes that if more ESE, district and school program staff set up programs that take into consideration the needs of evaluation design, we’ll more frequently be able to determine what worked and why. Each time we have the opportunity to evaluate a program we learn more about how to improve it and whether and how it was successful. Conducting evaluations put us that much closer to serving the educators and students of the Commonwealth as well as we possibly can.

You can track your progress as you complete each of the steps included in this document using the [Putting It All Together Checklist](#Putting). We encourage all staff members to use the assistance of the Office of Planning and Research and contact Kendra Winner at the **early stages of** **program planning**. However, it’s not too late to begin thinking about evaluation even if your program is underway. Among other things, Kendra can answer questions about and assist with creating logic models and help with any of the key elements of planning an evaluation. Kendra can be reached at kwinner@doe.mass.edu.

This guide is presented in two major sections:

1. Defining Your Program Using Logic Models
2. Planning Your Evaluation
3. **Defining Your Program Using Logic Models**

Planning and evaluation go hand in hand. The logic model is a useful tool for program planning and evaluation purposes. A logic model is a systematic way to visually outline your theory of the relationships among the resources you have to operate your program, the activities you plan to do, and the changes or impacts you hope to achieve. Whether it’s a program to get parents more involved in their child’s education or a program to improve the reading comprehension of middle schoolers, all programs[[1]](#footnote-1) have a goal to change or improve something. Logic models are commonly used to clarify the intentions and expectations of a program and can be useful for every phase of program development: design, implementation, and evaluation.

**Figure 1: Elements of a Logic Model**

| **Resources (Inputs)** | **Strategies and Activities** | **Outputs** | **Short-term outcomes** | **Long-term outcomes** | **Impacts** |
| --- | --- | --- | --- | --- | --- |

A program is more likely to succeed when there is agreement among stakeholders about the strategies and sequence of events that need to occur in order to realistically accomplish program goals and objectives.[[2]](#footnote-2) One application of a logic model is to get everyone on the same page about the program and the approach the program will take to create change. Stakeholders can have very different ideas about the purposes of a program, and the strategies to be used to facilitate the desired outcomes. These differences can surface during the planning phase of a program. Involving key stakeholders in the logic model development process helps ensure that stakeholders share a common understanding of the program. For these reasons, identifying the appropriate individuals who should be involved in drafting and reviewing the logic model is important.

The logic model process is used to clarify:

*During program planning*

* What the program will do.
* Who will participate in and benefit from the program activities.
* How the program will address unmet needs or existing gaps.
* How the program components (activities, outputs, and outcomes) logically fit together.

*During evaluation planning*

* How progress and success will be defined and measured.
* What will be evaluated; when the evaluation activities will take place; and who will be responsible for gathering, analyzing, and disseminating the findings.
* How lessons learned will be shared and used for ongoing improvement.

Working out a logic model also helps you to identify what you already know and what you don’t know about your program. For example, you may know that year-long residency models produce new teachers who are more likely to stay in the profession, but you may not know how districts can sustainably fund these models. Creating a logic model will also help to identify any gaps or assumptions in the overall program design that you might not have considered. Creating a logic model is also a way of doing your homework so that you don’t waste time developing research questions to which you already know the answer.

***What is a logic model?[[3]](#footnote-3)***

* A program-planning tool for organizing, designing, implementing, and evaluating any kind of program.
* A simplified picture of a program initiative.
* A model of the logical relationships among the resources that are invested, the activities that take place, and the benefits or changes that result from a program.

A logic model is NOT a fully developed project plan or evaluation design. More work is required to create both project and evaluation plans. For example, a logic model does not include specific deliverables, assignments, or research methodology. A logic model is useful in designing project plans and evaluation designs because it helps to clarify both what the program is and what it is not. A logic model can also help to identify areas of strength and/or weaknesses. This kind of clarification is helpful in building both a project plan and an evaluation design. As project plans and evaluation plans change, the logic model can be adjusted to reflect these plans.

Typically logic models have the following components:

* **Problem statement**: Defines the issue or problem that you are trying to address with your program.
* **Resources or inputs**: Specifies the money, materials, and equipment, people, time and partnerships needed.
* **Strategies and activities**: Describes what you propose to do to address the problem: activities, services, events, and products—*what your program does*.
* **Outputs**: Articulates the immediate results of the activities in the logic model—*what your program produces*. Outputs provide concrete, measurable evidence that the activity occurred. Outputs can be counted: number of teachers who participated, number of workshops held. Outputs don’t address change in behavior or change in how things are done.
* **Outcomes**: Outlines the difference the program makes in the short and long term—*what your program achieves*.
	+ **Short term**: most immediate and measurable results for participants that can be attributed to strategies and activities (e.g., teacher satisfaction, student interest).
	+ **Long term**: more distant, though anticipated, results of participation in strategies and activities (e.g., quality of STEM curricula, student learning outcomes).
* **Impact**: Specifies the far-reaching, ultimate influence you hope your program will have on the issue at hand (e.g., more middle and high school students pursue STEM degrees in college).
* **Assumptions**: beliefs about participants, staff, the program, and how change or improvement may be realized.

Because the differences between these different logic model components can sometimes be confusing, the logic model resources on page 4 offer more detail and information about logic models.

We recommend conducting the process of creating a logic model by combining two strategies: design forward from the known activities and design backward from the effects. We suggest starting first with moving backward from the effects. To do this, identify the desired long-term goals and definitions of success then work backward to identify all the conditions that must be in place for the goals to occur. This represents your “theory of action” about how change will occur. This backward mapping strategy facilitates the process of first identifying specific outcomes and/or definitions of success for the program. After the outcomes have first been identified, then you can designate the resources, strategies and short-term outcomes that promote the long-term success of your program. Conducting this step prevents you from just matching outcomes to the available resources.

The way in which you define success is essential to being able to determine how you can interpret your findings. Clear and specific definitions for success will help you have an unambiguous understanding of the results. Figure 2 compares how a specific definition of success can lead to more actionable interpretations of results.

**Figure 2: Defining Success**

| ***How would you know if the program was successful?*** |
| --- |
| **Unspecific Definition of Success:** | **Specific Definition of Success** |
| “Student’s STEM achievement will increase.” | “85% of participating MS & HS students’ learning outcomes in STEM improve” |
| **Program Results:**35% of participating MS & HS students’ learning outcomes in STEM improve |
| * Conversation will focus on “was the growth enough?”
* Stakeholders can interpret results as proof of success OR failure
* It may be difficult to take action due to a lack of consensus
 | * Conversation will focus on “Why did the program not meet the definition of success?”
* Stakeholders will have a common understanding of the results
* Districts can take action to fix, target, reduce, or eliminate programs based on the outcomes
 |

Table 1 on page 5 illustrates one common way a logic model can be depicted. This logic model will be referenced throughout the rest of the document to illustrate how a logic model contributes to designing and evaluating a program. A sample blank template can be found in Appendix A: [Blank Logic Model Template.](#Template)

***Links to Logic Model Resources***

[Difference Between Inputs, Activities, Outputs, Outcomes, and Impact](https://evaluateblog.wordpress.com/2013/06/10/difference-between-inputs-activities-outputs-outcomes-and-impact/)

[Theory of Change: A Practical Tool for Action, Results, and Learning](http://www.aecf.org/resources/theory-of-change/)

[Logic Models: A Tool For Designing and Monitoring Program Evaluations](https://ies.ed.gov/ncee/edlabs/regions/pacific/pdf/REL_2014007.pdf)

[Logic Models For Program Design, Implementation, and Evaluation: Workshop Toolkit](http://files.eric.ed.gov/fulltext/ED556231.pdf)

**Table 1: Sample STEM Program Logic Model**

| **Problem Statement:** Too few of the students in the district pursue STEM in higher education. |
| --- |
| **Resources (Inputs)** | **Strategies and Activities** | **Outputs** | **Short-term outcomes** | **Long-term outcomes** | **Impacts** |
| *Certain resources are needed to operate your program* | ***If*** *you have access to these resources,* ***then*** *you can use them to accomplish your planned activities* | ***If*** *you accomplish the stated activities,* ***then*** *you will produce the following amount of product and/or services:* | ***If*** *you accomplish these activities,* ***then*** *participants will benefit in the following ways in the short term:* | ***If*** *you accomplish these activities,* ***then*** *participants will benefit in the following ways in the long term:* | ***If*** *you accomplish these activities,* ***then*** *the following impacts or changes in systems will occur in the long term:* |
| District middle and high school educators and studentsUniversity facilities and staff: faculty and studentsLocal STEM businessesLocal museumsCommunity partners, including parents | District teachers attend summer workshops at university labs and participate in professional learning communities (PLCs) during the school yearUniversity faculty and students mentor teachers with inquiry-based science and lesson planning | Number of middle and high school teachers participatingNumber of summer workshops and PLC meetingsNumber of inquiry-based lessons createdNumber of university faculty and university students participating as mentors | 85% of teachers report satisfaction with professional development in inquiry-based STEM at the university85% of teachers use the professional development to create inquiry-based curriculum85% of PLC members have a positive experience | STEM PLCs become part of the schools’ routinesThe quality of STEM curricula in schools increasesAll participating teachers and university staff feel professionally connected to a broader STEM community of researchers and practitioners85% of participating MS & HS students’ learning outcomes (e.g., assessment results) in STEM improve | 25% more MS & HS students pursue STEM in higher education than prior to the program |
| **Assumptions:** 1) The professional development is of high quality; 2) The district and university leaders provide support for the STEM Partnership; 3) Teachers can apply what they have learned into inquiry-based curricula; 4) The PLCs are productive spaces for educator learning. |

# Planning Your Evaluation

Once you’ve completed a logic model for your program, you can move on to planning your evaluation. Program evaluation builds evidence about your program. It is the process of carefully collecting and analyzing information (or data) about a program or some aspect of a program that allows you to make necessary decisions about the program. The type of evaluation you undertake depends on what you want to learn about the program and the available resources and timeline for the evaluation. Most evaluations fall into one of two broad types: formative and summative.

A **formative evaluation** focuses on the implementation and processes of the program. Formative evaluation occurs from the beginning of a program and throughout the program and will allow you to identify mid-stream corrections or changes that need to be made. From Table 1, the logic model example on page 5, a formative evaluation looks at the strategies and activities as well as the outputs and outcomes of the program. A formative evaluation asks questions like: Does the program function the way it is meant to? What were the barriers or challenges?

In contrast, a **summative evaluation** focuses on the outcomes or impacts of a program; these are the short-term, long-term, and impacts portions of the logic model. Summative evaluations focus on whether and how the end goals of the evaluation, short- and long term-outcomes or, possibly, impacts, were achieved and thus tend to occur at the end of a program with data collected at key points in time to capture progress in the selected measures. A summative evaluation might ask questions like: Did specific student or educator outcomes improve as a result of the program? How much?

**Table 2: The Logic Model and Formative and Summative Evaluations**

| **Resources (Inputs)** | **Strategies and Activities** | **Outputs** | **Short-term outcomes** | **Long-term outcomes** | **Impacts** |
| --- | --- | --- | --- | --- | --- |
| **Formative Evaluation** |  |
|  | **Summative Evaluation** |

When possible, conduct both formative and summative evaluations. They can be part of the same overall evaluation process. Both evaluation methods are recommended for use simultaneously to provide program staff with ongoing feedback for program modifications (formative) as well as periodic review of long-term progress on major program goals and objectives (summative). Table 3 below outlines a comparison of formative and summative evaluations.

**Table 3: Comparison of Formative and Summative Evaluations.**

|  | **Formative** | **Summative** |
| --- | --- | --- |
| Typical Stage of Program | During program development or implementation | End of program |
| Types of Information | How best to revise and modify for improvement | Whether a program should be fixed, targeted, reduced or eliminated  |
| Helpful for | Ongoing feedback for program modifications | Review of long-term progress on major program goals and objectives |

**Prerequisites for High Quality Evaluations**

This next set of sections outline some major considerations to keep in mind as you are planning your program and/or evaluation in order to make it possible to use the most rigorous design for your evaluation. You can benefit from briefly reflecting on what you might get after working through this document. The steps outlined in this guide are designed to support your work though key elements of planning an evaluation:

* Define your program (the logic model)
* Draft actionable research questions
* Determine how to measure progress
* Decide on a research design and analysis
* Select whom to study (sample)
* Select data collection strategies
* Weigh the tradeoffs of an internal or an external evaluation
* Use your results

You can track your progress as you complete each of these steps using the **[Putting It All Together](#Putting)** checklist on page 16.

***Draft actionable research questions.***

Evaluations are organized around answering questions, which are typically called the research questions. The evaluation is a strategy designed to provide information to answer the question(s) of interest. Developing good research questions, like a good logic model, is a critical step in the evaluation process. The research questions, when appropriately written, will guide the project and assist in the construction of a logical argument. The research questions should be clear, focused questions and summarize the issue being investigated.[[4]](#footnote-4) To be measurable, questions need to be “operationalized.” Questions are operationalized when the meaning of each term in the research question can be defined. Table 4 provides an example of how common terms need to be clearly specified in order to be operationalized.

Examples of research questions from the STEM program case could include but are not limited to:

* Compared with before the program, how many STEM PLCs are in place after the program?
* Compared with before the program what percentage of program participants attend STEM PLCs after participating in the program?
* Compared with similar individuals who did not participate in the program, what percentage of teachers and university staff feel professionally connected to a broader STEM community of researchers and practitioners after participation in the program?

**Table 4: Operationalizing Research Questions**

| **Question**: How many teachers are participating in the STEM program?[[5]](#footnote-5) |
| --- |
| **Component** | **Issues to be clarified** |
| How many | Does the “how many” refer to a head count or full-time equivalent (FTE) count? |
| teachers | Do “teachers” include only certified teachers? Only teachers certified in a STEM area? Teachers teaching in STEM areas without certification? |
| Are | At what point in time should teachers be considered participants? At the beginning of the program? At the end? Only those who complete all of the program activities? |
| participating in  | Does this include teachers who only participate in some of the aspects of the intervention? Participate for only part of the time? Or teachers who complete all of the program activities over the entire course of the program? |
| the STEM program | What counts as the STEM program? The combination of all aspects of the program? All summer workshops, mentoring, PLC meetings AND creation of inquiry-based curriculum? |

Next are some considerations to help in the development of research questions that will yield useful information.

* *Are the questions guided by the logic model?* *Do they reinforce the logic model?* Logic models help to identify clear, specific, and actionable evaluation questions. Questions can be identified before the logic model is developed, but selecting specific evaluation questions usually comes after a logic model has been created and program staff members have a clear understanding of the evaluation resources needed to adequately address them.[[6]](#footnote-6)
* *Are the questions practical and commensurate with the capacity and timeline that you have to answer them?* For example, while the ultimate impact of the logic model above suggests the question “Do more middle school and high school students pursue STEM in higher education?” That question requires many years to address. An alternative question could be, “Do students who participated in the STEM program more frequently report an interest in STEM college courses (or majors) than those students who did not participate in the program?”

*Will answers to the questions yield actionable information? Will the answers help you make decisions? Improve, scale up, or scale back?* Think through what you are going to DO with the information from your questions. This step will help to narrow research questions and check that the responses to questions will provide you with actionable information. **Unless you can do something with the answers to your research questions, you’re wasting valuable resources.** For example, from the logic model, a short-term outcome of the program is for PLC members to have a positive experience. However, the research question should be asked in such a way that you can take action on the information that you gather. In addition to asking PLC members to rate the degree to which their experience is positive, you could also ask, “What do PLC members find useful about the PLC? What do they find challenging about the PLC?” The answers to these questions would allow you to identify PLC strengths that could be further supported and challenges that need to be addressed.

* *Are the questions measurable?* The ultimate impact of the program outlined on in Table 1 on page 5 suggests the following research question, “Do more middle school and high school students pursue STEM in higher education?” To answer this question there must be data that already exist or could easily be collected that measures this outcome. Defining “pursue STEM in higher education” is essential to identifying the correct data. The researchers could define this as students who declare the intention to major or minor in a STEM subject area. This data could be collected by conducting a survey of participating students as they approach high school graduation. However, if the researchers define this as only those students who actually take college course work or complete a degree in STEM, the data are not so easily attainable. The Massachusetts Department of Higher Education (DHE) has good data on post-secondary degrees, majors and courses but this information is limited to those students who attend Massachusetts public colleges. Depending on the size and scale of the project, the researchers could consider tracking students once they graduate from high school and survey them during their college experience or when they graduate from college. However, the researchers should take into account the extent to which they are likely to lose track of students over the intervening years. It is important to have a plan for how the data necessary will be available or gathered to answer the questions.

***Determine How to Measure Progress.***

Similarly, indicators of progress and/or success should be defined with enough specificity to make them measurable. Indicators are only useful to the degree to which they accurately reflect what they are intended to measure. Setting clear and measurable definitions of success for your program is necessary to interpreting your evaluation results.

The following provide a checklist of criteria to use in selecting appropriate indicators:[[7]](#footnote-7)

* **Relevance:** There is a clear relationship between the indicator and the program.
* **Accuracy:** The indicator measures what it purports to measure.
* **Importance:** The measurement captures something that makes a difference in program effectiveness.
* **Usefulness:** The indicator captures information that helps advance the program and the results point to areas for improvement.
* **Feasibility:** Data can be obtained with reasonable and affordable effort.
* **Credibility:** The indicator is being used by leading experts and organizations.
* **Validity:**The extent to which the measure actually reflects what it sets out to measure or the reality it claims torepresent.
* **Distinctiveness:** The indicator lacks redundancy and does not measure something already captured under other indicators.

When selecting indicators, consider long-term as well as short-term indicators and how each will be measured. Short-term indicators measure whether programs and initiatives are on track to reaching long-term goals. Short-term indicators, or indicators of early evidence of change, identify changes you should begin to see if the plan is having its desired impact while being implemented. These indicators might include changes in practice or attitude from sources such as classroom observation or surveys. For example, from our logic model, we might see that 75% of participating teachers improved their STEM curricula over an academic year.

***Decide on a research design and analysis.***

This section is presented to make readers aware of some of the basic elements of research design and analysis. This introductory information on comparison groups and causal designs is meant to acquaint you enough with these concepts so that you have some knowledge of what options are available to you. OPR hopes that this section makes it more likely that you will design a program so that such analyses could be carried out by an experienced evaluator. This section is not intended to support you in carrying out these analyses on your own.

A research design is a plan for how to answer the research questions. A research design lays out which methods are best used for answering which questions. A research design also maps out how each method will be utilized and articulates the limitations of each method.

In order to answer research questions, some research designs require a specific kind of comparison group be defined so that you can say what your results compare to. For instance, are you comparing your results to business as usual? To other similar students or teachers who did not participate in the program? Comparison groups determine the confidence with which you can answer certain types of research questions and define how your results “compare to what?” Below you will find four different ways to construct a meaningful comparison between students (or others, e.g., teachers, college faculty, college students).

* Comparison to similar students/teachers: Comparing outcomes of otherwise similar students who were randomly assigned to either participate in a program or not. This is the comparison sample required for methods that can demonstrate your program actually caused the outcomes you chose to document (see page 11 for more description of experimental and quasi-experimental methods).
* Comparison to baseline data: Comparing the relevant outcomes of students and/or teachers in a program to previous outcomes for the same students or the same teachers.
* Comparison to internal standards: Comparing the outcomes of students in the program to a district performance standard.
* Comparison to external benchmarks: Comparing the outcomes of students in a program to outcomes from a broader pool of students outside the district.

You could, for example, compare the teachers and students participating in the STEM program example to a number of other kinds of groups. You could decide to not to pick a comparison group and compare participating teachers and students to themselves prior to the program using pre- and post-testing (comparison to baseline data). In this example you would collect baseline data (data from before the program) on variables such as students’ learning outcomes and the number and quality of STEM PLCs prior to the program. However, you will not be able to draw any conclusions about causality with this comparison. If you want to know if the STEM program *causes* any of your outcomes of interest, you would use either an experimental design or a quasi-experimental design. For example, let’s assume that the STEM program is working with multiple schools. In this case you could randomly select two groups of schools, those that participate in the program and those that don’t participate in the program, and compare the two groups at the conclusion of the program. This method set you up to determine if any differences you see in the groups are caused by the program and not other factors.

While many quantitative research designs can answer your evaluation questions, they are not all designed to tell you whether something was caused by something else. There are different types of quantitative analysis which yield different types of information, some analyses are causal and some are not. Causal studies are designed to take advantage of randomization to prevent biases from being introduced into the study. For instance, to provide a simple example from our logic model, say we are testing the effectiveness of the STEM program on student learning outcomes. If we allow students from the class to volunteer to participate in the program, and we then compare the volunteers’ behavior against those who did not participate, our results will reflect something other than just the effects of the STEM program. Volunteers likely have characteristics that make them different from students who do not volunteer. In contrast, by randomly assigning subjects to be in the group that participates in the program (or “the treatment”) or to be in the control group, researchers can measure the effect of the program. However, random assignment can be difficult to implement in the real world – e.g., assigning students randomly to teachers or schools. For that reason, we present a hierarchy of designs you can aspire to with a brief description of each.

* + Experimental design. Whenever the evaluation focuses on outcomes, if it is possible, conduct a randomized control trial (RTC) as your quantitative research design. An RCT can be used when participants can be assigned randomly to either a treatment or non-treatment group. In the education context, this design is most likely to work well when a program is rolling out implementation over a number of years (staggered randomization) or when a lottery approach is used to assign participants to the treatment or non-treatment group. In a staggered randomization, the first group to get treated in, for example, year 1, is compared to a second untreated group in year 1 who will receive the treatment in year 2. Alternately, selecting participants based on a lottery system means that those applicants who are randomly selected to participate in a program (e.g., charter schools), should not differ from the group of applicants not selected to participate in the program as all applicants had an equal likelihood of being assigned to either group. Consider these methods of assignment whenever the agency cannot serve everyone who wants to participate (e.g., students, teachers, schools, etc.), as randomization is viewed as a way to give everyone a fair chance at access to the program and we can randomly select from among those qualified for the initiative.
	+ Quasi-Experimental designs. A quasi-experiment is a study used to estimate the causal impact of an intervention on its target population without random assignment. Two of the most commonly used quasi-experimental designs are regression discontinuity design (RDD) and comparative interrupted time series (CITS). An RDD relies on a clear cut point that is unknown to participants ahead of time (e.g. test scores, percentile rankings of schools or accountability ratings of districts, etc…). CITS relies on comparing the trends for two groups pre- and post-treatment by selecting comparison groups that have similar demographic characteristics. These kinds of research design are the next strongest level of evidence to a RCT. If an RCT cannot be used, these quasi-experimental designs are the next preferable choice for quantitative design.
	+ Comparative analyses. Even when causal designs are not feasible, side by side examination of alternatives such as sets of data from two or more alternative processes permits a meaningful comparative analysis. For example, comparisons can be made between programs that are yielding strong outcomes vs. those that are not, or variations on the same program, to help narrow in on the factors/conditions for success (e.g., comparing the student outcomes of teachers who participate in all aspects of the STEM program to those teachers who only participate in PLCs). While RTCs and quasi-experimental designs can tell us about causality, good descriptive quantitative data is still valuable and can lay the groundwork for identifying potential comparisons. Descriptive quantitative data can tell you a lot about how many and how much, for example how many summer workshops and PLC meetings participating teachers attend.

***Select whom to study.***

Generally, you can use one of three different ways to decide whom to study, each with advantages and challenges:

* Census: Includes everyone or everything in the population being studied – for example, all teachers, principals, or superintendents in Massachusetts. A census allows you to make generalizations about the population because the entire population is included. A census can be conducted if the entire population is very small or it is reasonable to include the entire population (e.g., a survey of all superintendents in Massachusetts). The benefit of this method is that you capture the entire group you are attempting to study. The challenge of this strategy is that it can be difficult to get enough participants from the entire population to participate in order to be a true census.
* Representative Sample: If you would like to be able to make generalizations about a particular group, and the population you are interested in is too large to attempt to include all of its members (e.g., all secondary school students in the state), select a [representative sample](https://www.investopedia.com/terms/r/representative-sample.asp). A representative sample is one that properly reflects the population from which it is chosen, for example, a sample of teachers that reflect all teachers in the state of Massachusetts. A representative sample is a subgroup of the entire population of interest that emulates the entire population. Selecting a representative sample can be complex as it has to be constructed to reasonably reflect the entire population it is meant to represent. In order to generate a representative teacher sample, for example, you have to take into account variables such as age, gender, ethnicity, length of tenure, and potentially other variables that will make your sample representative. The process of selecting a representative sample is not easy and best left to researchers who must use a battery of statistical techniques to ensure that the subset is as representative as possible.
* Convenience Sample: Chosen based on the accessibility of individuals and their willingness to participate in the research. This method can be fairly easy to implement. However, when using statistics to analyze trends, you cannot rely on convenience sampling to make statistical inferences about larger populations. Convenience sampling does not generate samples that adequately represent the population the samples are meant to reflect. Because of this, it is not possible to generalize the findings to a larger group.

***Select data collection strategies that reflect your research questions.***

On page 15, you will find a chart that illustrates data collection strategies and their relative advantages and challenges. The table provides an overview of the major methods used for collecting data during evaluations.

District staff members also have access to their own existing student demographic data (SIMS) and student achievement data (MCAS), among other. In addition, ESE has created aggregate school- and district-level files to make student assessment results (MCAS and PARCC), data used for accountability determinations, and higher education enrollment data from the National Student Clearinghouse available in a more accessible format to district and school staff. As these become available, ESE posts them at <http://www.doe.mass.edu/infoservices/research/>.

***Weigh the tradeoffs of conducting an internal or an external evaluation.***

Some evaluations or evaluation activities might make sense to conduct internally instead of externally and vice versa. In some cases, internal staff will be in a good position to carry out an evaluation as they are familiar with the actual administration of the program. At minimum, for example, an internal evaluation could conduct surveys with participating teachers and students. Checklists could also be collected of the activities teachers participate in (summer workshops and PLCs) as well as the STEM activities that students attend.

Even if budgets are tight and hiring an external evaluator is not possible, it’s still a good idea to do some sort of evaluation using internal staff. Table 5 outlines some general ways in which you can think about the tradeoffs of methods and the degree to which they are time and cost intensive as well as whether the information they gather tends to be broad or deep. This can guide you in thinking about what types of data collection you might tackle if you have to utilize internal staff in addition to or instead of an external evaluator.

**Table 5: Data Collection Method Tradeoffs**

|  | **Less Time and Cost Intensive** | **More Time and Cost Intensive** |
| --- | --- | --- |
| **Broad** | Questionnaires, surveys, checklists | Observations (if many are conducted) |
| **Deep** | Documentation review | Interviews, focus groups, case studies |

One major difference between an external evaluation and an internal evaluation is the extent to which internal staff can be and are perceived to be impartial. If a topic is politically sensitive or requires a study that will stand up to criticism, OPR recommends that an external, third party organization or individual conduct the evaluation if at all possible. For example, when ESE wanted an evaluation of the roll out of our educator evaluation system, we hired an external organization with experience in educator evaluation to address any claims that the results were influenced by ESE. This evaluation included surveying a representative sample of teachers and principals, as well as conducting case studies and focus groups of representative districts implementing the new system.

***Use your results.***

Once you’ve completed your evaluation, you’ll want to do something with your results. The ultimate goal of an evaluation is to take action on programs based on evidence of effectiveness. **Unless you can do something with the answers to your research questions, you’re wasting valuable resources.** Depending on how you’ve crafted your evaluation and your findings, you can use evaluation results in a variety of ways. For example, you could:

* **Expand or keep** programs that you’ve been able to determine cause increases in student achievement.
* **Target** particular groups of students or teachers for whom the program was successful and focusthe program on those populations.
* **Reduce** programs that have only some success and keep only the crucial portions.
* **Fix** programs where success would improve if particular problems were fixed.
* **Eliminate** programs that prove to be ineffective.

***Links to Evaluation Resources***

[Basic Guide to Program Evaluation](http://managementhelp.org/evaluation/program-evaluation-guide.htm)
[Measuring Program Outcomes: A Practical Approach](http://digitalcommons.unomaha.edu/cgi/viewcontent.cgi?article=1047&context=slceeval)

[Program Manager's Guide to Evaluation](http://www.acf.hhs.gov/programs/opre/other_resrch/pm_guide_eval/index.html)

[W.K. Kellogg Foundation Evaluation Handbook](http://www.wkkf.org/knowledge-center/resources/2010/W-K-Kellogg-Foundation-Evaluation-Handbook.aspx)

**Table 6. Overview of Methods for Data Collection[[8]](#footnote-8)**

| **Method** | **Overall Purpose** | **Advantages** | **Challenges** |
| --- | --- | --- | --- |
| **Questionnaires, surveys, checklists** | To quickly and/or easily gather lots of information from people in a non-threatening way | * Can be completed anonymously
* Inexpensive to administer
* Easy to compare and analyze
* Can be administered to many people
* Can get lots of data
* Many samples already exist
 | * Might not get thoughtful feedback
* Wording can bias participant responses
* Impersonal
* For surveys, may need a sampling expert
* Doesn’t get full story
 |
| **Document review** | To get an impression of how a program operates without interrupting the program; from review of applications, finances, memos, minutes, etc… | * Get comprehensive and historical information
* Doesn’t interrupt program or participants’ routines in program
* Information already exists
* Few biases introduced about information
 | * Can be time consuming
* Info make be incomplete
* Info may require clarification from program
* Need to be quite clear about what looking for
* Restricted to existing data
 |
| **Interviews** | To fully understand someone’s impressions or experiences, or learn more about their answers to questionnaires | * Get full range and depth of information
* Develop relationships with participants
* Can be flexible with participants
 | * Can be time intensive and costly
* Can be hard to analyze and compare
* Interview can bias participants’ responses
 |
| **Focus groups** | To explore a topic in depth through group discussion, e.g., about reactions to an experience or suggestion, understanding common complaints, etc… | * Quickly and reliably get common and disparate impressions
* Can be efficient way to get both breadth and depth of information in a short time
* Can convey key information about programs
 | * Participants can influence each other’s responses
* Can be hard to analyze responses
* Need good facilitator
* Difficult to schedule 6-8 people together
* Can be time consuming and expensive
 |
| **Observations** | To gather accurate information about how a program actually operates, particularly about processes | * View operations of a program as they are actually occurring
* Can adapt to events as they occur
 | * Can be difficult to interpret seen behaviors
* Can be complex to categorize observations
* Observer can influence behaviors of program participants
* Can be time consuming and expensive
 |
| **Case studies** | To fully understand or depict participants’ experiences in a program, and conduct comprehensive examination through cross comparison of cases | * Fully depicts participants’ experience in program input, process, and results
* Powerful means to portray program to outsiders
 | * Usually quite time consuming and expensive to collect, organize, and describe
* Represents depth of information rather than breadth
 |

**Figure 2. Putting It All Together Checklist**

| **Designing Your Program** | **Notes** |
| --- | --- |
|  | * Problem Statement
* Resources/Inputs
* Strategies/Activities
* Outputs
* Short-term Outcomes
* Long-term Outcomes
* Impact
* Assumptions
 |  |
| **Planning Your Evaluation** |  |
| **Research Questions** | * Operationalized
* Guided by logic model
* Practical
* Actionable
* Measurable
 |  |
| **Measures of Progress** | * Relevant
* Accurate
* Important
* Useful
* Feasible
* Credible
* Valid
* Distinctive
 |  |
| **Research Design and Analysis** | * Randomized Control Trial (RCT)
* Quasi-experimental (CITS, RDD)
* Comparative Analysis
* Other
 |  |
| **Comparison Group** | * Similar Teachers/Students
* Baseline data
* Internal Standards
* External Benchmarks
 |  |
| **Sample** | * Census
* Representative Sample
* Convenience Sample
 |  |
| **Data Collection Strategies** | * Questionnaires/Surveys/Checklists
* Document Review
* Interviews
* Focus Group
* Observation
* Case Studies
* Existing Data Sets
 |  |
| **Internal/External Evaluation** | * Internal Evaluation
* External Evaluation
* Combination
 |  |
| **Use Your Findings** | * Expand
* Target
* Reduce
* Fix
* Eliminate
 |  |

**Appendix A:** **[Blank Logic Model Template](%5C%5C%5C%5CESE-FPS-MAL-002.doe.mass.edu%5C%5CHOME%5C%5CKLW%5C%5CESSA%20Evidence%20Based%20Practices%5C%5CEvaluation%20%20Guide%5C%5C2018%20January%20versions%5C%5CTemplate)**

| **Problem Statement:**  |
| --- |
| **Resources (Inputs)** | **Strategies and Activities** | **Outputs** | **Short-term outcomes** | **Long-term outcomes** | **Impacts** |
| *Certain resources are needed to operate your program* | ***If*** *you have access to these resources,* ***then*** *you can use them to accomplish your planned activities* | ***If*** *you accomplish the stated activities,* ***then*** *you will produce the following amount of product and/or services:* | ***If*** *you accomplish these activities,* ***then*** *participants will benefit in the following ways in the short term:* | ***If*** *you accomplish these activities,* ***then*** *participants will benefit in the following ways in the long term:* | ***If*** *you accomplish these activities,* ***then*** *the following impacts or changes in systems will occur in the long term:* |
|  |  |  |  |  |  |
| **Assumptions:**  |

1. Throughout the document the word “program” refers to programs, interventions or activities [↑](#footnote-ref-1)
2. Portions of this taken from Silverman, B., Mai, C., Boulet, S., and O’Leary, L. *Logic Models for Planning and Evaluation: A Resource Guide for the CDC State Birth Defects Surveillance Program Cooperative Agreement*. Retrieved on January 1, 2018 from <https://www.cdc.gov/ncbddd/birthdefects/models/resource1-evaluationguide2009.pdf> [↑](#footnote-ref-2)
3. Portions of this section taken from University of Wisconsin-Extension (2003). *Enhancing Program Performance with Logic Models*. University of Wisconsin-Extension. Retrieved on January 5, 2018 from <https://fyi.uwex.edu/programdevelopment/files/2016/03/lmcourseall.pdf> [↑](#footnote-ref-3)
4. For more information on developing research questions see Center for Innovation in Research and Teaching. *Writing a Good Research Question*. Retrieved on January 5, 2018 from <https://cirt.gcu.edu/research/developmentresources/tutorials/question> [↑](#footnote-ref-4)
5. Adapted from National Forum on Education Statistics. (2014). *Forum Guide to Supporting Data Access for Researchers: A Local Education Agency Perspective*. (NFES 2014-801). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved on January 5 2018 from <https://nces.ed.gov/pubs2014/2014801.pdf> [↑](#footnote-ref-5)
6. Lawton, B., Brandon, P.R., Cicchinelli, L., & Kekahio, W. (2014). *Logic models: A tool for designing and monitoring program evaluations.* (REL 2014–007). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Region-al Assistance, Regional Educational Laboratory Pacific. Retrieved on January 5, 2018 from <https://ies.ed.gov/ncee/edlabs/regions/pacific/pdf/REL_2014007.pdf> [↑](#footnote-ref-6)
7. Adapted from MacDonald, G. *Criteria for Selection of High-Performing Indicators: A Checklist to Inform Monitoring and Evaluation. Centers for Disease Control and Prevention*. Atlanta, Georgia. Retrieved on January 5, 2018 from <http://www.betterevaluation.org/sites/default/files/Indicator_checklist.pdf> [↑](#footnote-ref-7)
8. #  Adapted from C. McNamara. *Overview of Basic Methods to Collect Information*. Retrieved on January 5, 2018 from <https://managementhelp.org/businessresearch/methods.htm>

 [↑](#footnote-ref-8)