WestEd
WestEd.org

Magnet School Evaluation Guide

Amanda Nabors, Jonathan Nakamoto,
Rachel Tripathy, Sara Allender

Suggested citation: Nabors, A., Nakamoto, J., Tripathy, R., & Allender, S. (2023). *Magnet school evaluation guide*. WestEd.

WestEd is a nonpartisan, nonprofit research, development, and service agency that works with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide. More information about WestEd is available at WestEd.org.

Acknowledgments

The authors are thankful for the thoughtful reviews, suggestions, and contributions to this guide by their WestEd colleagues: Kimkinyona Cully, Arena Lam, Jason Snipes, Thomas Torrey-Gibney, and Juan Carlos Bojorquez.

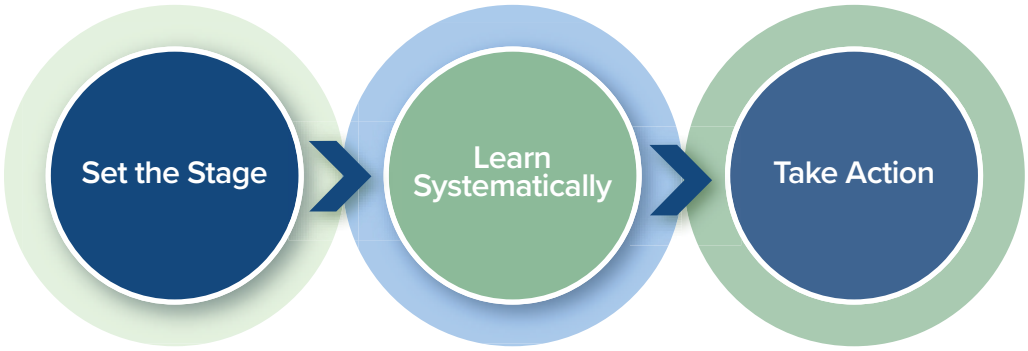
Contents

WestEd
WestEd.org

WestEd's Magnet School Evaluation Guide

This guide offers resources to help you facilitate an effective evaluation of your magnet program, whether the program is funded by Federal Magnet Schools Assistance Program (MSAP) grants or other sources of funding. For MSAP grantees, there are additional recommendations in this guide to help you self-assess your current evaluation plan or guide you in selection of an evaluation partner.

The guide will help you *set the stage* and prepare for the evaluation, *learn systematically* from the evaluation's findings, and then *take action* to improve your magnet programs. The guide answers commonly asked questions about evaluations and provides resources to help you work effectively with evaluators throughout the life of your magnet programs. It walks through practices and norms you can adopt during the planning and early grant phases that will set the stage for and strengthen your eventual evaluation. It describes the systematic phases that any evaluation should move through and provides examples of what that evaluation may look like and ways to ensure the evaluation is rigorous. Finally, the guide covers how to successfully learn from your evaluation, improve the implementation of your magnet program, and equitably disseminate the evaluation findings.



Set the Stage

This section offers foundational information about evaluations and describes important considerations to address in preparing for an evaluation of your magnet program.

What Is Evaluation?

Evaluation is the use of social science research methods to systematically assess programs that are designed to improve society (Rossi et al., 2019). For example, evaluations can help determine how, why, and under what circumstances magnet programs improve students' educational experiences and outcomes. Evaluations of programs and policies, such as an evaluation of a district magnet program, can be designed in many different ways depending on context and purpose.

A ***formative*** evaluation is designed to co-occur with program implementation activities such that evaluation findings can inform programmatic improvements in a timely and responsive manner. These evaluations typically include a feedback loop between evaluators and program leaders, such as school or district staff, wherein evaluation findings may be shared, co-interpreted, and applied in consultation with implementers in the field in a relatively rapid manner. For an evaluation of a science, technology, engineering, and mathematics (STEM) magnet program, for example, a formative evaluation could include assessing the professional development focused on project-based learning (PBL) for STEM teachers. Surveys and focus groups could be used for gauging the quality and effectiveness of the professional development and identifying areas for improvement. Findings that are reported rapidly to program leaders can guide modifications to the professional development before future rounds of trainings.

A ***summative*** evaluation typically describes the progress or impact achieved by a program at the conclusion of a specified implementation period—for example, once the original funding vehicle for the program has ended or once the intended level of program uptake has been achieved. These evaluations assess the success of program implementation activities, including any observed impacts on the target populations. For an evaluation of a performing arts magnet school serving students in grades 9–12, for example, a summative evaluation could examine the number of students who participated in each of the school's different programs (e.g., dance and theatre arts) and the number of graduates of the program who enrolled in 2- and 4-year colleges during the MSAP grant.

Some evaluations focus exclusively on program impacts, while others include a blend of formative and summative

components depending on when in the life cycle of a project the evaluation is occurring and the types of questions an evaluation is trying to answer. Some evaluations employ only quantitative methods—that is, focusing on the numbers—while others rely on only qualitative data, such as interviews and focus groups. Most MSAP-funded evaluations are mixed-method studies, meaning they employ several evaluation strategies at varying levels of rigor that utilize both quantitative and qualitative data.

An ***impact*** evaluation usually refers to an evaluation strategy that employs an experimental or quasi-experimental study to examine whether a program affected the outcomes of interest. Conducting an impact evaluation is important for accountability purposes and is a required component of an MSAP grant. Experimental studies, which are also called randomized controlled trials (RCTs), rely on randomization to create groups that participate or do not participate in the intervention. A magnet school's enrollment lottery is an example of a way to randomize participation and can be used as the basis for an RCT. Quasi-experimental studies generally involve the identification of matched comparison groups that are not participating in the intervention, and these studies use prior achievement and student demographics in the matching process. For an evaluation of a magnet school, the quasi-experimental design (QED) could include students who attended a traditional public school in the comparison group. Experimental and quasi-experimental studies can be challenging to execute but can provide strong evidence that, for example, enrollment in a specific magnet program caused improvements in student achievement.

Reporting evaluation findings can be an important first step in changing a policy, practice, or process because evaluations can identify areas for improvement and may include recommendations as part of synthesizing their results. Evaluations, and evaluators, can shed light—from a new perspective—on the ways in which a magnet program is working and for whom, but evaluators will never have the experience of being in a program or scenario every day. While evaluators are often happy to partner on thinking about solutions to issues surfaced by an evaluation, you are the experts of your own magnet programs.

WHAT IS A RIGOROUS IMPACT EVALUATION?

A research design for an impact evaluation that is statistically rigorous gives you more confidence in your ability to detect program effects, whether positive or negative, and thereby to accurately assess the true impact of your program. Doing so allows you to confidently identify program successes that you could scale or areas for improvement to address. Additionally, meeting defined thresholds of statistical rigor is increasingly important to competitions for federal grant funds. For example, the MSAP requires that grantees contract with independent, external evaluators to design a study that will meet at least a level of evidence standards known as "promising"—though higher tiers of evidence allow you to draw conclusions about program effectiveness that promising evidence cannot.

The U.S. Department of Education's definition of "promising" evidence is "empirical evidence to support the theoretical linkage(s) between at least one critical component and at least one relevant outcome presented in the logic model for the proposed process, product, strategy, or practice" (Stoker, 2022, p. 28). To provide that empirical evidence—to be able to say that your program is causing a particular impact—the study design must be rigorous. For the MSAP, this expectation means that a study design must be either

- a quasi-experimental study or
- an RCT.

(The MSAP grant application guidelines additionally highlight "correlational studies" as a potential design to produce "promising evidence." Correlational studies are weaker QEDs that do not identify a comparison group that is equivalent to the intervention group on critical baseline measures such as prior achievement and are not likely to allow an application to receive all seven points for the impact evaluation portion of the scoring criteria. As a result, this guide focuses only on QEDs and RCTs.)

The importance of a rigorous impact evaluation is highlighted by the most recent MSAP grant application (fiscal year 2022), which allocated 7 of the 20 evaluation points to whether the grantee's proposed plan would likely produce "promising evidence" about the impact of the magnet programming on a student outcome.

More detail on different possible study designs is provided later in the guide. More on evidence of promise and evidence tiers can be found at the following:

- U.S. Department of Education's slide deck "Quality of Project Evaluation: Producing Evidence of Promise"
- What Works Clearinghouse's Evidence Tiers and WWC Ratings Resources

Prepare for an Evaluation

Before you can begin an evaluation, taking several preparatory steps can help ensure you get the most out of the evaluation.

Choose the right evaluation partner

While some school districts have evaluators as part of their organizations, it can also be helpful to get an outside perspective. External evaluators can work with you in a variety of capacities, from providing full-service evaluations to being content specialists who assist an in-house evaluation to serving as advisors and thought partners. Once you have chosen an evaluation partner, you can work collaboratively with them to identify the goals of the evaluation, research questions, and research designs.

Evaluators often have different specialties, from content areas to methodological approaches. An evaluator who will best suit your needs likely has experience doing work in your particular environment and context (e.g., urban, rural, STEM, etc.), an evaluation portfolio that includes other work with school districts and state agencies, and a methodological approach that is well suited to meeting the goals of your evaluation. If you are seeking to conduct a rigorous evaluation, you will want to look for an evaluation partner with experience conducting RCTs and quasi-experimental studies. Smaller evaluation firms may be able to provide lower cost evaluations. On the other hand, larger evaluation firms will likely have individuals who specialize in each of the potential methodologies that might be needed for your evaluation and can provide a full "menu" of different services.

Identify goals

Your evaluation partner can help your internal team determine goals for the proposed evaluation. These do not need to be formal research questions; rather, your team should establish broadly what they want to learn from the evaluation. Your internal team and evaluation partner can help facilitate meetings with key collaborators, additional partners, and other pertinent audiences to translate their thinking into evaluation goals. Key collaborators and partners can include district leaders, school staff, professional development providers, students, and parents, including those who share the same background and/or experiences as the program participants. Involving key collaborators and partners from the beginning and throughout the evaluation process is key to ensuring the accuracy of the study findings and trust in the evaluation process (WestEd, 2021). The evaluators can help your internal team and collaborators and partners operationalize the goals during the evaluation process.

Identify key personnel

Team composition is key when choosing an evaluation partner and when building your internal team. Evaluation teams that are representative of the program participants, whether by background or experience, are better situated to understand participant experiences and the historical, political, and cultural context in which the program is situated (WestEd, 2021). While composing your internal team, consider which personnel will need to be connected most closely to your evaluation partners. These personnel should be staff who know your intervention and your local context, such as district-level program staff, key collaborators and partners, supervisors, and data managers, as well as school-based staff.

QUESTIONS TO CONSIDER WHEN DEVELOPING EVALUATION GOALS

- What is the nature of the intervention/policy/program?
    - Is the purpose or focus new for your context?
    - What implementation features are new to your context? What implementation features are not new?
    - What resources are necessary to implement the intervention?
- Whom is the intervention intended to impact?
    - Which students? Which teachers or faculty? Which facilities? Which families/community members?
    - How might those impacted groups be involved in the evaluation?
    - What challenges or biases exist that may inhibit support, acceptance, or implementation of the intervention?
- Who is carrying out the intervention?
    - Which teachers or faculty? Which schools/facilities?
- What are the desired impacts/changes/outcomes?
    - What might success look like for each impacted group? In the short term? In the long term?
    - What changes in behaviors, beliefs, performance, conditions, or other indicators would indicate success?
    - What unintended consequences might result from this intervention?

Establish working norms

You and your evaluation partner will benefit from establishing a set of shared working norms or meeting agreements. Setting up ongoing communication channels and rhythms, establishing meeting protocols, and cocreating timelines for the work will aid the success of your evaluation. Part of establishing shared norms is identifying and addressing how implicit biases as well as overt racism, sexism, and other forms of oppression may influence the evaluation. By engaging in thoughtful reflections and conversations up front, your evaluation team can strengthen its members' shared awareness and understandings of project priorities and local contexts (WestEd, 2021).

Learn Systematically

This section dives into the details and considerations that are important for actually creating and implementing an effective evaluation of your magnet program.

Develop a Theory of Change

A *theory of change* is a high-level "North Star" description of how a program will lead to its intended goals. There are many ways to conceptualize, design, and create a theory of change (see, e.g., The Annie E. Casey Foundation, 2022); in other resources, you may see "theory of change" and "logic model" used interchangeably. For the purposes of this guide, a theory of change is operationalized as a summary statement that concisely describes the key program activities, outputs, and outcomes and works in tandem with a logic model to visually display and further detail the approach. After you establish the baseline purpose of the program or intervention, there are two approaches that you and collaborators can follow to build a theory of change. You can start by listing out your resources or program activities and then work *forward* to what you intend to achieve, or you can start with the impacts you wish to see and work *backward* to think about the resources that can get you there. The process of crafting a theory of change should clarify the pathway(s) through which change is expected to happen.

THEORY OF CHANGE EXAMPLE SUMMARY STATEMENTS

"By doing [program activities], participants will [action that results in changed skills, behavior, or experience], leading to [desired outcome]."

"By centering project-based learning in every course, our program will connect the students' school experience to real-world applications of STEM, increasing their capacity for inquiry-driven learning and strengthening their preparation for college and careers."

As you develop each section of your theory of change, ask yourself the following questions:

- What assumptions am I making about how this program is supposed to work?
  - What are my assumptions about how implementation of activities will occur?
  - What are my assumptions about participants or other collaborators and partners?
  - How might historical contexts, systems of oppression, or bias play into those assumptions?
- What factors outside of the program's control may affect key program activities?
  - How might local priorities or interests evolve over the course of implementation?

SYSTEMIC OPPRESSION AND ACKNOWLEDGING CONTEXT

As you consider the assumptions that underlie your program and theory of change, be clear about the root causes that have led to the need for the program (WestEd, 2021), and consider who has the power and impetus to make the change you want to see. Being explicit in your discussions about the ways in which participating groups are affected by the pressures of a larger system of oppression—including both how those systems affect their day-to-day experiences and impact their access to social, community, and material resources, etc.—will help you more accurately identify the root causes of the problem you wish to impact. Knowing the root causes in turn will help you create a more robust, effective theory of change.

For example, if your program aims to increase the share of Latina girls reading at grade level in the 3rd grade, you might initially phrase your statement of the problem as "Latina girls in 3rd grade are behind in reading." That conceptualization puts the onus for change on the children and not the systems that are poorly serving them. A problem statement that situates the issue within fuller context might be phrased as "Our school is not preparing Latina girls to read at grade level by the 3rd grade."

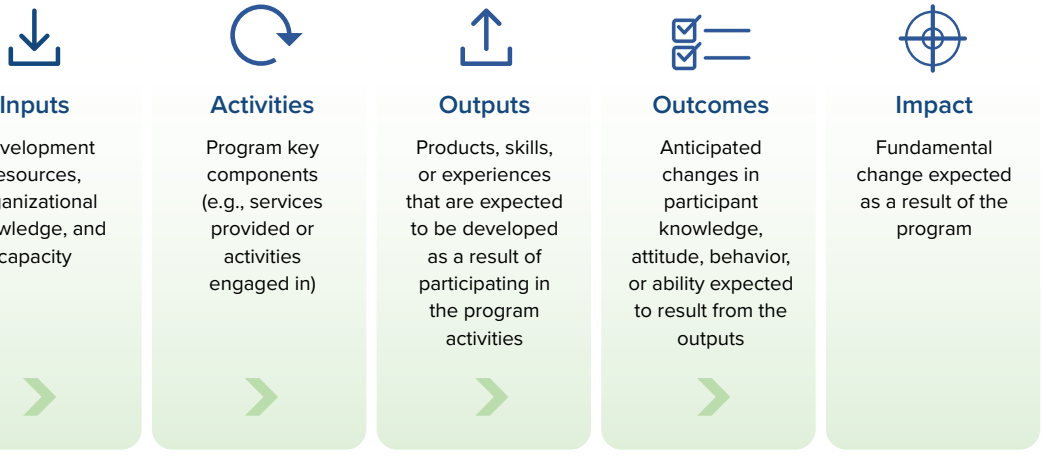
Develop a Logic Model

A *logic model* is a visual answer to the question, "How will your magnet program improve student, school, and community outcomes?" It elaborates on your theory of change in a graphic format by showing how different program activities (what you do) contribute to your program goals (what will happen). A well-crafted logic model will give you a systematic framework for identifying the key **inputs**, **activities**, **outputs**, **outcomes**, **and impacts** of your program, which will in turn provide clarity on the appropriate research questions, measures, and measurement instruments for the evaluation.

Creating a logic model is a great way to ensure the entire team shares an understanding about the goals of the magnet program and the planned path to achieving those goals. This visual guide can also serve as a useful resource for your team and for participants, funders, and partners who may enter the project later. The logic model is a living resource and may need to be adjusted as you learn more about the realities of implementing the magnet program.

A logic model can be read as a series of causal "if-then" statements that describe how your program will achieve its intended outcomes: "If [activities] occur, then [outputs] will happen, ultimately leading to [outcomes] and [impact]." Using the following Logic Model Example Template, you can build a logic model from your theory of change. From the logic model, you will be well situated to develop strong evaluation questions.

Logic Model Example Template



- Inputs: What are the key resources that will make up or contribute to the program or intervention?
- Activities: What actions will happen as part of the program that are expected to lead to the desired change?
- Outputs: What is expected to be created through the program activities? These can be material (a policy handbook) or immaterial (increased skill in supporting students' social–emotional well-being).
- Outcomes: What changes are the outputs meant to lead to? Outcomes can often be categorized into short-term, intermediate, and end-of-program outcomes.
- Impact: What are the long-term implications of the outcomes of the program? What changes do you expect to see after the program is over?

ADDITIONAL LOGIC MODEL RESOURCES

Here are some additional templates, guides, and toolkits to help build your logic model:

- Developing Logic Models is a slide deck released by WestEd's National Center for Systemic Improvement (NCSI).
- Logic Models for Program Design, Implementation, and Evaluation: Workshop Toolkit includes a facilitator workbook, participant workbook, and slide deck created by the Regional Educational Laboratory (REL) Northeast & Islands.
- Logic Model Planning Worksheet is a tool that features group activities and three templates from the Doing What Works Library.
- Tearless Logic Model is a resource from the *Global Journal of Community Psychology Practice*.

Build an Evaluation Plan

Once you have established a theory of change and a logic model, you are ready to identify evaluation questions and design an evaluation that can answer them as rigorously as possible. Each of the arrows in your logic model represents a hypothesized causal link that can be tested. Do you want to know if the activities are leading to the anticipated outputs or if the outputs are leading to the intended outcomes? If you have an outside funder for the program, they may have key questions or required outcomes to report on. Your evaluation partner can serve as a thought partner to develop or refine these questions with you depending on the needs of your program. Evaluation questions and designs should be relevant, reasonable, and rigorous.

- **Relevant:** Evaluation questions should be relevant to the purpose and outcomes of your context and program. They should be appropriate to the type of evaluation you are performing. **Process-focused questions** are appropriate for evaluating the arrows between inputs, activities, and outputs. **Outcome-focused questions** are appropriate for evaluating the arrows between outputs, outcomes, and impacts.
- **Reasonable:** While asking challenging evaluation questions that require rigorous designs to answer is encouraged, the questions need to be answerable in your context. If you do not have the expertise, time, and funding to perform an RCT, your evaluation questions should not require such an evaluation to be answered.
- **Rigorous:** While you should be reasonable about the limitations of your team and context, strive to ask the most rigorous questions you can within that context. A question being methodologically difficult to answer is not the same as it being practically difficult to answer. The rigor of your evaluation questions determines the rigor of your evaluation design, which affects the conclusions you can draw from the evaluation results.

SAMPLE EVALUATION QUESTIONS

PROCESS

To what extent are magnet program activities implemented as intended? What factors facilitate or hinder implementation?

What are the perceptions of teachers, staff, and school leaders about the implementation and effectiveness of the magnet program?

Is the program reaching the intended population? Is there evidence of variation in program uptake or implementation exposure among certain groups?

OUTCOME

To what extent does attendance at a magnet school increase student achievement in mathematics and English language arts?

To what extent do observed impacts vary by student characteristics, including race/ethnicity, gender, socioeconomic status, English Learner status, or participation in special education?

To what extent is the magnet program reducing minority group isolation in program schools?

Once you have identified your evaluation questions, it is time to design a **rigorous** evaluation that can answer them in concert with your evaluation partner. Choosing the right measurements is as important as choosing what to measure. One way to ensure the usefulness and validity of your evaluation questions and design is to cocreate them with groups representative of the participants of the program as well as program partners or have them reviewed by those groups and partners. The evaluation should be designed in such a way as to minimize bias, such as by implementing designs informed by the work of a diverse set of scholars; including recruitment and sampling strategies that accurately represent the population being studied; including culturally appropriate methods for gathering data; considering culturally appropriate qualitative and quantitative instruments, tools, and data collection processes; and disaggregating data by race and ethnicity as defined by collaborators and partners (WestEd, 2021). Including these considerations can increase community buy-in for the evaluation, ensure evaluation findings are relevant to various partners, and ensure that the findings' implications are actionable.

Gather High-Quality Data

While you plan your evaluation, your team can begin to assess your data and data system needs, plan for additional data collection (if necessary), and make informed decisions about evaluation structure and design based on available data. While evaluations often include some form of data collection, the starting point should always be extant data—that is, data that are already available within your system. It may be helpful to ask questions such as the following:

- What sources of data do you have access to now? How are they stored? At what level are they aggregated—that is, are they collected at the student level, the teacher level, the school level, or some other level?
- How trustworthy are the data? How recent are the data? How are the data collected?
- Who collected the data? Who analyzed the data? What biases and beliefs underlie the collection and analysis? What gaps exist in the data?
- Who has access to key pieces of data?
- What do project partners and participants consider useful and important data pieces?
- What data are you missing? What gaps in data collection or availability currently prevent you from answering key questions about your program?

POSSIBLE DATA SOURCES

Data that may aid the evaluation come in many forms. There is no one correct set of data that you must have for a successful evaluation. Depending on the research design and project goals, your evaluator will ask you to gather extant data from your school or district and may want to collect additional data throughout the course of the evaluation. The following are some examples of commonly used types of data in magnet program evaluations.

QUANTITATIVE DATA

Student demographics

Community demographics

Student performance data

School performance data

Surveys with Likert scale items

Class registrar data

QUALITATIVE DATA

Administrator interviews

Teacher focus groups

Classroom walk-throughs or observations

Surveys with open-ended questions

As you develop a sense of the gaps and strengths of your extant data, you will be better positioned to plan for data collection as part of the evaluation. High-quality and timely data collection requires intentional work ahead of time. As part of your evaluation design, your evaluation partner will aid you in developing a timeline and plan for communicating with key partners and participants about data collection, performing the collection, and conducting any follow-up necessary that fits within the larger evaluation and program implementation timeline. Plan early and check in often on data collection timelines. Considerations may include the following:

- Institutional Review Boards (IRBs): Part of the process of ensuring the ethical allowability of, protection of people involved in, and quality of an evaluation can be to submit evaluation plans to an IRB. Such an application may take some time to be completed, reviewed, and approved.
- Data Sharing Agreements (DSAs) or Memorandums of Understanding (MOUs): DSAs and MOUs help all parties involved in an evaluation understand what data are being used and collected; how they are being used; and who is responsible for the collection, use, and protection of those data. These documents may require sign-off from data privacy offices, legal offices, and other departments. Developing these documents and collecting the signatures should be incorporated into the data collection timeline.
- Data Security Plans: Such plans are documents that define the data used for evaluations and how the data will be managed. Data security plans are particularly important for evaluations that rely on student-level data and data with direct and indirect identifiers. The purpose of a data security plan is to identify and describe the strategies used to protect, store, and access the data used in an evaluation. Once documented, the plan can be used to assess whether actual practice follows what was documented and can serve as a way to communicate to project team members what the standard security practices are.

Evaluate the Implementation of a Magnet Program

Evaluating implementation (i.e., examining fidelity of implementation; Century et al., 2010) is an often-overlooked part of evaluations. If you do not fully understand how a magnet program is being implemented, you cannot completely understand the outputs and outcomes. Schools and districts are busy places doing many things at once, and it is likely that not every aspect of a magnet program will go according to plan. Understanding the differences between the proposed activities and the actual activities that are implemented will help you interpret the results from an impact evaluation. For example, smaller impacts on student outcomes would be easier to justify if the implementation evaluation indicated the magnet program was not implemented fully. The implementation evaluation will also allow you to attribute outcomes to the appropriate factors and know what to do differently or what to replicate in the future.

The following are some questions to consider when evaluating implementation:

- What are the key components of the magnet program?
- What are the best ways to measure how the key components are being implemented?
- Is the magnet program being implemented as planned?
- What would you consider to be low, medium, and high levels of implementation of each of the key components?

An important first step when examining the implementation of a magnet program is to identify the critical components of the program (Century et al., 2010). The critical components may include concrete activities such as providing teacher professional development and offering student clubs related to the magnet theme. The critical components may also include quality of delivery measures that, for example, assess the effectiveness of the professional development and students' participation levels and enthusiasm for the student clubs (Dane & Schneider, 1998). The critical components of a magnet program can be identified through discussions with school and district staff, a review of the logic model, and a review of magnet program documents.

Once the critical components of a magnet program are identified, it is important to develop a measurement strategy for each of the components. In most magnet school evaluations, a combination of program records (e.g., professional development attendance sheets), surveys of staff and students, and interviews with district or school staff can be used to assess the implementation of each component.

It is also important for the evaluator, while developing the measurement strategies, to work collaboratively with magnet staff to identify thresholds for low, medium, and high levels of implementation for each

component. Creating and agreeing upon these thresholds can be time-consuming but will allow the evaluation report to include concrete and easy-to-understand descriptions of the overall level of implementation of a magnet program and the level of implementation of each individual component of the program. Additionally, in some situations, it may be possible to examine impacts on students who participated to a great extent in magnet programming (i.e., they experienced a program with high fidelity to the program model), which would provide an upper bound on possible magnet program impacts.

Examples of a Magnet Program's Critical Components and Thresholds for Levels of Implementation

Critical component	Sub-component	Measurement strategy	Low	Medium	High
Teacher professional development (PD)	Hours of PD completed	Attendance records	Teachers completed an average of less than 60% of PD hours	Teachers completed an average of 60–90% of PD hours	Teachers completed an average of 90–100% of PD hours
Teacher PD	Effectiveness of the PD	Teacher survey that uses a 1 (*not at all effective*) to 4 (*very effective*) rating scale	Teachers rated that the PD was below 2.5 on average	Teachers rated that the PD was between 2.5 and 3.5 on average	Teachers rated that the PD was 3.5 or higher on average

Evaluate Outcomes and Impacts of a Magnet Program

The goal of an impact evaluation is to determine whether a causal relationship exists between, for example, attendance at a magnet school and improved mathematics achievement. In this example, the critical question is whether the students would have shown the same level of mathematics achievement had they not attended the magnet school. Assessing the impact of a magnet program using a rigorous design will be a critical part of most evaluation plans. As noted above, the scoring criteria for the most recent MSAP grant application allocates one third of the evaluation points to whether the grantee's proposed plan would likely produce evidence about the impact of the magnet programming on a student outcome.

Your evaluation plan and logic model will guide how you evaluate your magnet program's outcomes and impacts. Magnet school evaluations will generally examine impacts on student outcomes, such as achievement on standardized tests, graduation rates, or attendance. However, the evaluations could also examine the impact of the magnet programs on other outcomes, such as the use of specific teaching practices. Rigorously evaluating the impact of your magnet program on the outcomes in your logic model will require careful measurement strategies and in-depth knowledge of research design, database management, and statistical analysis.

Choosing measures

An important part of an impact evaluation is to choose the measures used to assess the outcomes of interest. Many student outcomes—including grade point average (GPA), graduation rates, attendance rates, and discipline incidents—can be obtained directly from school and district records. Measuring other outcomes requires robust instruments, such as surveys and assessments, that take time to develop and must be carefully designed and administered to ensure that they collect valid and reliable data. That is, researchers need to ensure that the instruments are measuring what they propose to measure and are doing so consistently. The U.S. Department of Education's What Works Clearinghouse (WWC) has reporting standards for validity and reliability, which should be adhered to in order to give credibility to an evaluation (What Works Clearinghouse, 2022). Additionally, MSAP grants currently require impact studies that adhere to the WWC standards.

One way to ensure that the measures you use have sufficient validity and reliability is to select instruments that are widely used and have already been tested numerous times. There is often no need to start from scratch. Standardized tests, such as the Smarter Balanced assessments and the ACT, already have documented validity and reliability. When you are developing instruments with your evaluation partner, it may be important to consider the best practices for survey development (Dillman et

al., 2014) or educational testing (American Educational Research Association, 2014). Reflections on Applying Principles of Equitable Evaluation (Stern et al., 2019), from WestEd's Justice and Prevention Research Center, outlines several ways to test validity and reliability. Beyond traditional calculations of validity and reliability, any measures that are developed for your study should also be reviewed by partners and other collaborators and by partners who are representative of the subjects of your magnet program. The ways in which instruments are structured, worded, used, and analyzed are not value-neutral, and a review by diverse individuals is a key part of conducting equitable evaluations and can lead to higher quality data collections (Stern et al., 2019).

Considerations of rigorous designs: Defining randomized controlled trials and quasi-experimental designs

RCTs and QEDs are the two types of rigorous designs that are commonly used when evaluating the impact of magnet programs. RCTs, which rely on randomly assigning individuals or groups to conditions, are more rigorous than QEDs and allow for stronger causal conclusions about the impact of a magnet program. QEDs, or observational studies, contrast the outcomes of those who received an intervention with the outcomes of those who did not participate in the intervention. Although QEDs are not as rigorous as RCTs, causal conclusions about the impact of a magnet program can be drawn from a QED when the comparison group is equivalent to the intervention group prior to the start of the intervention.

RCT: An RCT is a research design in which individuals or groups (e.g., schools) are randomly assigned to a treatment group that receives the intervention of interest (e.g., a magnet program) and to a control group that receives an alternative treatment (typically, a business-as-usual condition such as attending a traditional public school). When randomization is successful, the two groups are similar on observed and unobserved variables prior to the start of the treatment. Outcomes for the treatment group are contrasted against outcomes from the control group to determine the impact of the intervention.

QED: There are several types of QEDs, but the most rigorous type is a design in which outcomes for the treatment group (i.e., individuals or groups receiving the intervention of interest) are contrasted with the outcomes for a comparison group. Rather than being identified via random assignment, the comparison group is identified from a preexisting group that did not participate in the intervention and is matched to the treatment group so that the groups are statistically equivalent prior to the start of the intervention. For magnet school evaluations, the most critical variables to match on are measures of prior achievement and student demographics.

QEDs cannot account for unobserved differences between the groups on factors such as student motivation, which may impact student achievement.

RCTs are implemented in education research in a variety of ways. For example, researchers can randomly assign students to participate or not participate in an after-school tutoring program, or researchers can randomly assign entire schools to participate in a new professional development program or the business-as-usual professional development. Within the context of magnet schools, an "opportunistic experiment" (Resch et al., 2014) based on the magnet schools' lotteries is the most common way that an RCT is used. This type of RCT relies on the randomization process in the schools' lotteries that identify which students attend the schools. In this design, outcomes for students who won a school's lottery and gained admittance to that school are contrasted with the outcomes for students who did not win the school's lottery and therefore could not attend the magnet school. There are a number of complexities and requirements to successfully employing this type of design, most notably the need for a school to have many more students applying to the school than it can accept. As a result, it is always good practice to consider whether an RCT is possible for an evaluation; however, when an RCT is not viable, the next best option is a QED.

Consistent with RCTs, QEDs are implemented in a multitude of ways in education research. For example, when a new schoolwide teacher professional development program focusing on mathematics instruction is rolled out to a small number of schools in a district, researchers can identify a pool of comparison schools not participating in the program from the pool of all nonparticipating schools in the district. For this example, the comparison schools could be matched on prior school-level mathematics achievement and student demographics so that they are equivalent at baseline to the program schools. The matching is usually done by what is known as Mahalanobis distance matching (Stuart, 2010), a multivariate matching algorithm that works well with small sample sizes. For magnet school evaluations, a common approach to doing a QED is to use the students attending a magnet school as the intervention group and then identify comparison students from a pool of other students in the same district who are not attending a magnet school. The comparison students would be identified using propensity score matching (i.e., another multivariate matching algorithm that works well with larger samples; Guo & Fraser, 2010) and matched on a number of variables, such as prior achievement and demographic characteristics.

For both RCTs and QEDs, it is important to plan for the database management requirements that will be needed to create the final data sets for the analyses. For a magnet

school evaluation examining impacts on student achievement, it is likely that data sets that include the students' test scores, demographics, and school enrollment will need to be combined using unique student identification numbers (or scrambled student identification numbers if required by an MOU or the IRB). These data sets also need to be merged across multiple years for tracking cohorts of students across time. Magnet school evaluations could also require student survey data to be merged with other data sources and could require the use of course records or attendance records. These data sets may include many thousands of students if the study is a QED and the pool of potential comparison students is large.

There are many different statistical analyses that researchers commonly use for RCTs and QEDs in education research. One example is hierarchical linear modeling, which appropriately accounts for the structure of education data, with students nested in classrooms and schools (Raudenbush & Bryk, 2002). For magnet school evaluations that examine the impact of a single magnet school on student achievement, multiple regression is a common approach used to analyze the data, as it allows researchers to statistically control for factors that may differ between treatment and comparison groups, such as prior achievement and demographics. However, other methods are also allowed by the WWC, such as the difference-in-difference analysis.[1] To meet the WWC standards, which would allow a magnet school evaluation to produce strong evidence regarding the impact of a magnet program, there are other analytic issues that need to be considered (What Works Clearinghouse, 2022), including how missing data will be handled, the calculation of attrition rates (for RCTs only), baseline equivalence (i.e., whether the groups are equivalent prior to the start of the intervention), and the calculation of effects sizes (i.e., a standardized way to calculate the size of a magnet school's impact; Lipsey et al., 2012). It is also critical to consider where your analytical methods may be introducing bias (e.g., by aggregating data by race and ethnicity categories that do not belong together).

1 The difference-in-difference analysis is based on subtracting the preintervention outcome mean from the postintervention outcome mean separately for the treatment and control groups and then subtracting the treatment group difference from the control group difference.

Potential RCT and QED Scenarios and Questions to Consider

Type of design	Potential use scenario	Important questions
Randomized controlled trial (RCT)	An evaluation of a magnet school with a lottery that has a large group of students (e.g., 100 or more students) who applied to the school and won the lottery and an equally large group of students who applied to the school and did not win the lottery in a given year	• What proportion of students who won the lottery decided to enroll at the school? • What proportion of students who did not win the lottery decided to attend a traditional public school or another magnet school? • Are baseline and outcome data available for the lottery winners and the lottery losers at approximately the same rates?
Quasi-experimental design (QED)	An evaluation of a magnet school that has a large group of students attending the school (e.g., 100 or more students) and has an even larger pool of potential comparison students (e.g., 500 or more students) from the same grade levels and from within the same district as the magnet school	• Are the comparison students likely to be similar to the magnet school students in terms of prior achievement and demographics? • Are the comparison students attending the same schools that the magnet school students would have attended if they had not enrolled at the magnet school?

Take Action

This section focuses on what to do with the findings of an evaluation and describes important considerations for *after* an evaluation has been completed.

Disseminate Findings

Throughout the course of your evaluation and at the end of your evaluation, you may have a number of dissemination obligations to a funder, such as the U.S. Department of Education; to district leadership; or to your collaborators and partners who helped implement your magnet program or design the evaluation. Evaluation reporting can follow the model of a traditional report or take the form of a data dashboard, an infographic, a series of briefs, a presentation, or a combination of several options. It is best practice to disseminate your results in a number of ways so that they are accessible to more groups because not everyone receives or processes information the same way. Additionally, it can be helpful to have collaborators and partners help to plan and implement the chosen dissemination strategies.

There is a substantial body of work outlining best practices related to disseminating evaluation findings. For example, methodology books are devoted to communicating and reporting evaluation findings (see Torres et al., 2005) and teaching how to create effective data visualizations (see Evergreen, 2020). Some of the best practices outlined by these authors include utilizing an organizing framework (e.g., an evaluation's research questions) to tailor reports for the intended audience and avoiding jargon and highly technical terms in evaluation reports (Torres et al., 2005). For data visualization, Evergreen (2020) presents a range of different types of graphs, describes the situations for which each type is best suited, and explains how to customize the graphs so that they can most effectively tell a story with the data.

When you are deciding what findings you plan to disseminate, some questions to ask include the following (WestEd, 2021):

- Who is this communication for, and what are their priorities?
- How does this audience prefer to be communicated with?
- What findings are most relevant to this audience?
- Is this product accessible for people who use screen readers or individuals with varying reading levels?
- Are the visuals created in a way that values the experiences of all participants and is culturally sensitive?
- Are subgroup analyses based on race and ethnicity being conducted appropriately?
- Are differences by race, gender, or socioeconomic status masked by the ways that the evaluation's data are aggregated or displayed?

- Has there been any "cherry-picking" of results or suppressing of negative findings?
- Are limitations of the measures or the analytic plan clearly described?

Validate Findings

Too often, communities are evaluated without an opportunity to provide feedback on their experiences as participants, weigh in on the findings, or even know what the findings are. This situation can lead to incomplete or inaccurate evaluation findings that can impact program improvement decisions and future funding opportunities (WestEd, 2021). Accordingly, before finalizing reporting, it is important to share back initial findings with project partners and, when possible, with participants. Having the involved parties validate the findings can strengthen your conclusions and help uncover gaps or implicit biases that emerged during data collection and/or analysis.

Make Recommendations

The final step of an evaluation is moving from findings to action. Recommendations can be some of the most important takeaways from an evaluation (Torres et al., 2005). Making data-based recommendations is one way to frame the findings and maintain momentum on projects. Having various collaborators and partners interpret the evaluation data can help you understand how the interpretations from the data impact the community and help you develop high-quality recommendations. Evaluation recommendations should be context- and project-specific and should be concrete, relevant, and reasonable. It may be helpful to organize your recommendations by evaluation time frame (e.g., actions that could be done immediately or actions that will require a longer timeline), the party the recommendation calls to action (e.g., magnet coordinator, principal, or district staff), or cost (e.g., actions that require no additional funding or actions that will require financial support). Finally, it can be very fruitful to work with your evaluators and other partners to create action plans so that the evaluation's recommendations are implemented (Torres et al., 2005).

SAMPLE RECOMMENDATIONS FROM PRIOR MAGNET SCHOOL EVALUATIONS

- Allocate time during or after school so that every teacher at the magnet school can work with the PBL coaches to help ensure that PBL activities are consistently implemented as a core instructional practice that is integral to the school's theme.
- Make sure that all students feel a strong connection to the school (particularly younger students who started school during the COVID-19 pandemic) through the use of schoolwide morale initiatives and strategic engagement for students in younger grades, specifically.
- Intentionally build in more time during the school day for students to work on their capstone PBL assignments with their groups to ensure that all students have completed assignments at the end of the year.
- Monitor messaging and communication related to the district's magnet school themes so that any disconnects can be easily identified.
- School administration should identify concrete goals for the coming school year related to the number of cross-curricular units to be developed and the number of subjects per unit.
- Develop systems or processes to support application of new learning. In other words, after teachers have engaged in a professional development opportunity, they should receive support in translating learnings from professional development into the classroom.
- Develop consistent and formalized needs-sensing related to professional development and coaching. This process could include teachers identifying their own needs; however, it should not be fully reliant on self-identification of needs. Develop a systematic approach to identify teachers' needs and to provide appropriate supports.

References

American Educational Research Association. (2014). *Standards for educational and psychological testing* (2014 ed.).

The Annie E. Casey Foundation. (2022). *Developing a theory of change: Practical theory of change guidance, templates, and examples*. https://www.aecf.org/resources/theory-of-change

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*(2), 199–218.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23–45.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The Tailored Design Method* (4th ed.). John Wiley and Sons.

Evergreen, S. D. H. (2020). *Effective data visualization: The right chart for the right data* (2nd ed.). Sage.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Sage.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Resch, A., Berk, J., & Akers, L. (2014). *Recognizing and conducting opportunistic experiments in education: A guide for policymakers and researchers* (REL 2014–037). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2019). *Evaluation: A systematic approach* (8th ed.). Sage.

Stern, A., Guckenburg, S., Persson, H., & Petrosino, A. (2019). *Reflections on applying principles of equitable evaluation*. WestEd. https://jprc.wested.org/project/reflections-on-applying-principles-of-equitable-evaluation/

Stoker, G. (2022). *Fiscal year 2022 pre-application webinar: Quality of project evaluation: Producing evidence of promise* [PowerPoint slides]. U.S. Department of Education. https://oese.ed.gov/files/2022/03EvidenceOfPromisePreappWebinarPresentation.pdf

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Torres, R. T., Preskill, H., & Piontek, M. E. (2005). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations*. Sage.

WestEd. (2021). *Anti-racist evaluation strategies: A guide for evaluation teams*. https://www.wested.org/resources/anti-racist-evaluation-strategies/

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Handbooks