Program Evaluation Toolkit Module 3, Chapter 2: Threats to Validity

Regional Educational Laboratory Central

From the National Center for Education Evaluation at IES

Speaker 1:

Welcome to the second chapter of Module 3. In this chapter, you will learn about threats to validity in evaluation design.

The purpose of program evaluation is to provide answers to questions about a program. Anyone can make unsupported claims about a program, but through the process of systematic evaluation and the intentional use of rigorous evaluation designs, you can answer your questions with confidence.

Evaluators refer to the concept of validity when discussing how confident they are in the claims of an evaluation. Validity refers to the extent to which the results of an evaluation are supportable, given the methods used. Validity applies to the evaluation design, analytic methods, and data collection. Ultimately, valid claims are sound ones.

You should examine threats to validity—factors that might undermine the soundness of your claims—regardless of what your evaluation design is or whether you ask process or outcome evaluation questions. Minimizing threats to validity is an essential part of any evaluation.

This chapter will provide you with a basic understanding about threats to validity so that you can be better equipped to understand and use evaluation. Additionally, this chapter will help you make sounder interpretations of evaluation results with an awareness of their limitations.

There are two main types of validity: internal validity and external validity. In this chapter, you will learn the difference between the two types and see examples of each. Both can be threatened in many ways, but this chapter focuses on only the most common ways.

First, let's look at *internal validity*, which relates to the results of an evaluation. At its most basic, internal validity involves asking whether the claims about the characteristics or effectiveness of a program are "true." Internal validity can be threatened by problems in the design, such as not controlling for key variables in a correlational or quasi-experimental design, or by problems in data collection or analysis, such as not properly accounting for missing data.

There are many threats to internal validity in program evaluation, but two of the most common threats are attrition and selection bias. Let's review these two common threats to internal validity.

Attrition happens when participants leave an evaluation after it begins but before it concludes. Attrition hinders an understanding of the outcomes of a program because there are often no data

on participants who leave, producing an incomplete picture of the impact of the program on the leavers.

Let's consider the AMMP! example. Remember that AMMP! is a fictitious after-school program to improve student math achievement and graduation rates. An evaluation of AMMP! might, among other things, estimate the program's effect on student dropout rates. But what if 15 percent of the students who were AMMP! participants at the start of the evaluation moved to other districts before graduation and were not included in final analyses? This 15 percent attrition rate would threaten the internal validity of the evaluation. Research has shown that highly mobile students are more likely to drop out of school. Therefore, if the evaluation finds that AMMP! has an effect on dropout prevention but that 15 percent of students in AMMP! moved to different districts, didn't finish the program, and were not included in the final analysis, the evaluation team should be concerned that this effect has been overestimated.

Selection bias happens when the treatment group differs from the comparison group in a meaningful way that is related to the outcomes of interest.

Again, let's consider this threat in relation to the AMMP! example. Say that a different district rolls out a program similar to AMMP! and allows any of its six middle schools to participate. Three schools chose to implement the program, and the district wants to conduct an evaluation to compare the outcomes for those three schools to the outcomes for the three schools not implementing the program.

This case represents a threat to internal validity due to selection bias. Meaningful differences between the treatment and comparison schools are likely. For instance, the schools that participate may be more likely to have staff that support program implementation, or they may serve student populations that receive more family support to complete homework assignments. As a result, the treatment group may not be representative of the target population, possibly leading to overly positive results.

Now let's discuss *external validity*, or generalizability. External validity is the extent to which the results of an evaluation may be generalized to different contexts, such as other populations or organizations. An evaluation can have strong internal validity but not have external validity. However, an evaluation must have internal validity in order to have external validity.

Next, let's look at contextual factors, multiple treatments, and the Hawthorne effect, as common threats to external validity.

Perhaps the most common way in which an evaluation lacks external validity, or generalizability, is different contextual factors between the sample in the evaluation and a population to which one wants to generalize. Consider again the AMMP! example. Let's say that AMMP! was implemented 15 years ago in a large urban district in the northeast United States. What if leaders in a rural district wished to implement AMMP! today? The leaders should be very careful in generalizing conclusions from the original AMMP! to the rural district. Because the time and place are so different, the strategies that were implemented effectively in the original context may not work in the second context.

Multiple treatments is also a threat to external validity, whereby external factors in addition to the program may cause the evaluation to detect a different effect than it would if the external factors were not present. For instance, perhaps AMMP! is being implemented in a district where a new math curriculum is also being implemented. In this example, AMMP! is the program being evaluated and a new math curriculum is an external factor not associated with AMMP!. Thus, a combination of the middle school's efforts to implement AMMP! and the new math curriculum could lead to effects that are not reproducible by either treatment alone. If a school district implements AMMP! in the future, that district could see a much different effect, since they did not have the same external factors.

Another example is called the Hawthorne effect, whereby individuals act differently because they know they are taking part in an evaluation. Sometimes, being studied can influence outcomes. What if AMMP! were a high-profile program with very visible funders and media coverage? This could lead to increased scrutiny, pressure, and praise—all of which might lead school staff and students to perform better than they would otherwise. Ultimately, the high-profile nature of the program could cause it to appear more effective than it would be without the visible funders and media coverage. In other words, some of the observed effects would be due to factors other than the evaluation of the actual program. If AMMP! were implemented later with less fanfare, it might lead to less positive outcomes. This is an example of the Hawthorne effect.

Many aspects of an evaluation may limit its validity. This chapter has provided a brief overview of some ways in which validity may be threatened. For now, keep in mind that you can assess the validity of an evaluation in terms of internal validity—or the extent to which results are justified based on the evidence gathered, including any relevant claims of cause and effect—and external validity—or the extent to which results may be generalized to other contexts.

Next, in chapter 3, you will learn about evidence of effectiveness and design standards.