Effect Size and Evaluation: The Basics

An impact evaluation aims to measure the impact of a program or policy on a priority outcome, and to detect a measurable change or effect. To plan for an evaluation, we need to decide how large or small an effect we want to be able to detect. This important decision will influence all aspects of evaluation planning, including budget, operations, duration, and sample. Considerations of effect size are crucial in designing an evaluation, but also rarely straightforward.

Consider the countless factors that determine the outcomes of a program or policy. For example, any outcome that depends on people's decisions or behavior — such as vaccination record, test scores, cholesterol level, or energy use — is influenced by numerous factors including attitudes, beliefs, motivations, life circumstances, personal experiences, and so on. A program or intervention aims to add onto those factors and change a specific outcome in a measurable way, and we often overestimate the effect that the program or intervention can actually achieve

What is an effect size?

Simply put, an effect size is the *magnitude* of the impact of a program or intervention on some outcome. For example, if the program is designed to promote vaccine uptake, then the effect size might be measured as a percentage-point increase in individuals who get a vaccine due to the specific program or intervention. If the program is designed to increase student achievement, then the effect size might be measured as an increase in average test scores. To ensure an evaluation is designed in such a way that you can detect an informative change, effect sizes need to be considered as part of an evaluation plan.

How does effect size matter in designing an evaluation?

Evaluators often refer to "detecting" an effect. Many people, on hearing this, may wonder, "If an effect is simply the average difference between two groups, then why can't we simply subtract the average in one group from the average in the other?" But imagine looking at two sticks, one which is exactly 12 inches long, and one which is 12 1/16 inches long. If you were to look at them both from, say, 5 feet away, you probably wouldn't be able to see any difference, and you might say that they were the same length. But if you looked at them up close, and especially if you looked at them with a magnifying glass, the difference would be obvious. Effect sizes are a little like the difference between the two sticks. When differences are very small, the statistical tools we have may not be powerful enough to "see," or detect, these differences. Designing an evaluation is like choosing a magnifying glass. To see smaller differences between sticks, you need a more powerful magnifying glass; to detect smaller effects of a program or intervention, you need a more powerful evaluation.

Every impact evaluation has a level of precision with which it can detect changes in an outcome. For example, an evaluation might have the power to detect a 3.5 percentage-point increase in vaccine uptake, or a 10 percentage-point increase in average test scores. "Power" is, in fact, the term that evaluators and statisticians use for this, and an evaluation's power depends on many factors, including sample size, base rate, variation in the outcome, randomization strategy, and background information that can be taken into account about the individuals or cases in the evaluation. Usually the most important, and most often within our control, is sample size. All else equal, a larger sample size gives us greater power and enables us to measure smaller effects with precision. The smallest effect that an evaluation can reliably detect is called a minimum detectable effect, or MDE.

Of course, greater sample size usually requires more time and money, which is why it is so important to design an evaluation around the effect size that we think we need to be able to measure. If we design our evaluation around an underestimated effect size, we may have a larger study than is required to answer our question or yield an actionable result. And, more common, if we design an evaluation around too large an expected effect size, we may design an underpowered study that is not capable of answering our question. Then, we may end up with an uninformative "null result."

Some common questions about effect sizes in evaluations

How do I know how small an effect my evaluation should be able to detect?

The answer to this question will depend on the program or intervention being evaluated and the needs of decision makers. One common approach is to try to predict the effect size that your program or intervention is likely to achieve and use this as the basis for your evaluation design. Unfortunately, it is often difficult to predict the effect size of a program or intervention in advance of actually running the evaluation. A second common approach is to ask, "What is the smallest effect that would justify adopting or scaling up the program or intervention?" The evaluation can then be designed with power to detect this effect size. If the evaluation detects no statistically reliable effect, then it is likely that any real effect would not have been policy relevant anyway.

Doesn't published literature give an accurate view of effect sizes?

Unfortunately, this is rarely the case. First, there is the now well-documented problem of publication bias. This means that results which show positive effects or larger effects are more likely to be published than evaluations that show no or small effects. The result is that published literature gives us an exaggerated picture of true effect sizes. This can cause us to overestimate the likely effect size and design an evaluation that can't detect the smaller real effects of our program. Second, there are few policy or programmatic areas which have comprehensive evidence published in the academic literature. Most areas may have some initial evidence or a set of relevant research results, but rarely is there a rich enough evidence base to make precise, confident predictions about likely effect sizes. Still, depending on the available research on programs or interventions like yours, you might be able to identify a plausible range of effect sizes that can serve as a rough basis for designing an evaluation.

Are small effect sizes actionable?

Small effects can be actionable. What is actionable depends on the policy or program. How expensive is the program or intervention? What are the alternatives that resources could be put toward instead? An expensive program or intervention may need to produce a large effect to hit its targets, whereas an inexpensive program or intervention may be useful or deemed effective even if only producing small effects.

Other Resources

Coppock, A. 10 Things to Know About Statistical Power. Methods guide published by Evidence in Governance and Politics (EGAP).

Kraft, M. A. (2019). *Interpreting Effect Sizes of Education Interventions.* EdWorkingPaper 19-10, Annenberg Institute at Brown University.

Innovations for Poverty Action. (2015). *Evaluating Financial Products. and Services in the US: A Toolkit for Running Randomized Controlled Trials.*