U.S. Department of Education
October 2016

What does it mean when a study finds no effect?

Neil Seftor
Mathematica Policy Research

REL 2017–265

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased, large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

October 2016

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Seftor, N. (2016). *What does it mean when a study finds no effects?* (REL 2017–265). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs.

Suppose you work for a school or district and need to make a decision about a program. For instance, maybe you are thinking about switching curricula or implementing a new schoolwide intervention aimed at behavior. You pick up a study of the program and read that there is no effect. What could have caused that result, and what does it mean? This brief explains what to look for in a study that finds no effect and how you can think about what to do next.

To measure program effectiveness, research studies compare outcomes for a group that received a program to outcomes of a group that did not. When a difference is not likely due to chance alone (that is, it is statistically significant), we say that the program had an effect.

But what do we conclude when desired differences are discovered, but are too small to be statistically significant? In these cases, we say there is no effect (see Box A).

Box A. No Effect Defined

In research studies, determining whether there is an effect is the result of a statistical test.

Researchers calculate the probability that an observed effect or difference could have occurred by chance alone.

If that probability is low enough, the observed effect or difference was very likely a real effect (it is statistically significant), and the researchers conclude that the program had an effect or impact.

Otherwise, there is not enough confidence that the observed effect or difference wasn't just due to chance. In these cases, there is said to be no effect.

This is a statistical definition only. In practice, a statistically significant finding may not be large enough to be meaningful. In other instances, a finding that is not statically significant may still be substantively important for practitioners or policy makers.

Three types of factors may contribute to a finding of no effect. The effects may be too small to be significant due to a failure of theory or a misunderstanding of how the theory should be turned into an intervention. Or perhaps what was supposed to be tested did not really happen—a failure of implementation. Or maybe the program had effects, but the study's design could not measure them with precision—a failure of research design. In science, we learn by getting things right and by getting things wrong. In both cases, it is important to dig deeper to understand what happened.

Example. Your district is considering switching to *Math Rocks* and decides to conduct a study. It involves 10 elementary schools, each with 4 third-grade classes of 25 students. In each school, half of the third-grade students are assigned to use *Math Rocks*. The rest continue to use *Math Is Awesome*. At the end of the year, students who used *Math Rocks* scored 8 points higher on the state assessment. However, the researchers conducting the study report that the 8-point difference is not statistically significant. What should you conclude about this finding that *Math Rocks* had no effect?

Situations like this are common. Consider the following findings from large-scale education evaluations and systematic reviews:

- The evaluation of the 21st Century Community Learning Centers Program found that those in the program scored no better on reading tests than those not in the program and had similar grades in English, mathematics, science, and social studies (James-Burdumy et al., 2005).
- In a study of 10 educational technology products used for reading and math instruction, differences in student test scores were not statistically significant between classrooms that used products and ones that did not (Dynarski et al., 2007).
- In the school year following participation in a supplemental reading program, there were no impacts on student academic performance (Somers et al., 2010).
- At the end of the second year of implementation, a development program for middle school math professionals did not have an impact on teacher knowledge of math or on student achievement (Garet et al., 2011).
- In systematic reviews of interventions designed to improve the reading comprehension of beginning readers, the What Works Clearinghouse™ found no effect for 13 interventions whose research met standards.

These findings gave information to the field that it did not have before the studies were done. All findings contribute to knowledge, including findings of no effect. But asking what else is learned from findings of no effect or how these findings can be used is useful.

A note about wording. The phrase *no effect* can be misinterpreted to mean that a program has *zero* effect on outcomes. But no effect does not mean zero effect. The program may improve outcomes at the same rate or extent as the program to which it was being compared. A more appropriate interpretation of no effect is that the program did not have a *detectably larger* effect on outcomes than the program to which it was being compared. Not having larger effects on outcomes is part of the iterative discovery process of identifying new ideas and approaches that work better. Not all ideas will pan out, but some will. Venture capitalists know that many new ideas or products will not be successful. But some will be as successful as Google or Facebook. Similarly, not all education innovations or programs will be effective, but some will be.

Is there really *no effect*?

Suppose a study reports no effect. Let's return to the three possible explanations—the theory failed, implementation failed, or the research design failed. Or some combination may account for the finding. Since failure of theory depends on eliminating failure of implementation and of research design, the latter two possibilities are discussed first.

How does a reader judge whether implementation failed? Ideally, the researchers gathered information and presented evidence of how well the program was implemented. The challenge is that few programs are specified so exactly that failure of implementation is obvious. Even if specifications are exact, studies may not have documented the quality of implementation. For

example, a program might call for teachers to have students on computers one to three times a week—a wide range. If students average one day a week on computers, is that reasonable implementation? Maybe. Or, for *Math Rocks*, suppose its designers indicate that teachers should get 20 hours of training on how to use it. The study reports that teachers got 10 hours of training. Is that failed implementation? Ten hours might be enough for experienced teachers and too little for inexperienced ones. Are staff experienced? Asking these kinds of questions helps a reader understand implementation context when the study does not address the success or failure of implementation directly.

How does a reader judge whether the research design failed? Failures of research design occur when the study's design could not measure effects of the program with precision or find meaningful effects to be statistically significant (see Box B). One of the most reliable guidelines in all of statistics is that studies with larger samples will estimate effects with greater precision. But researchers may have limited resources to carry out a study. For example, a study may have resources to include only 500 students rather than the 1,000 students needed to measure effects precisely. In this case, a finding of no effect may mean that the study was too small to detect a real effect.

Even if a study uses a large sample, the comparison group may differ from the program group for reasons unrelated to the program. This lack of equivalence between the groups prior to the start of the program could affect the estimate of the program's impact. Equivalence is less likely to be an issue for studies that rely on random assignment to form the groups being compared or account for baseline characteristics of the groups in the analysis.

Box B. Confidence Intervals

A confidence interval consists of a range of values surrounding the impact estimate. On the basis of the given data, a researcher using a valid design can be 95% confident that the true effect is between these bounds.

In the figure above, the circle is the impact estimate, the brackets are the bounds of the confidence interval, and the vertical line represents zero. If zero is within the interval, the finding is not statistically significant, and there is no effect. Otherwise, there is an effect.

The impact of two types of failures described in the text can be illustrated with this figure. A failure of implementation will move the point and brackets to the left, possibly enough that zero falls within the interval. A failure of research design may widen the brackets, again to possibly include zero in the interval. In both cases, a failure may result in a finding of no effect.

How does a reader judge whether theory failed? By the process of elimination: if the program was implemented as it was supposed to be implemented, and the research design was sound, then a finding of no effect means that a program had no net effect that was detected by the study. Sometimes a program that was expected to work actually doesn't, because the theory is flawed or wrong, or the theory was incorrectly turned into an intervention.

Returning to the example, readers of a study reporting that *Math Rocks* has no effect could reach different conclusions: *Math Rocks* is actually no better than *Math Is Awesome* (the theory suggesting

it is better is not sound and the no effect finding is correct), or *Math Rocks* really *is* better than *Math Is Awesome* (the theory is sound but either implementation or the research design failed).

Studies may intend to examine a well-implemented program using a sound design, but things may go awry. Asking a series of questions helps to distinguish failures of implementation or failures of research design. Table 1 shows the questions. Was there "fidelity" of implementation? Did students and teachers use the program to which they were assigned? Was the program being studied actually different from what it was being compared to? Were measures of outcomes the right ones? Did the study have enough power to detect an effect? Were the groups being compared in the study equivalent except for receipt of the program?

Table 1. How implementation and research design might contribute to a finding of no effect

Issue	Example	Questions to Ask Yourself
Fidelity of implementation. The way that a program is delivered may differ from how it was intended.	Some teachers exclude program components, shorten program sessions, or change the program delivery because of insufficient training, lack of motivation, or limited time.	• Were all of the program components implemented? • Were the components completed as planned? • Were personnel trained appropriately to implement the program? • Was the required support provided throughout the implementation? • Were personnel motivated to implement the program? • Was there enough time to implement the program (during each session and across the number of sessions)? • If there were implementation problems, can they be fixed? Or is it not feasible to implement the program as designed?
Noncompliance. Individuals do not comply with their original assignment to a curriculum or program.	Some teachers assigned to use *Math Rocks* may choose not to use it, or teachers assigned to *Math Is Awesome* may use *Math Rocks* instead.	• Did some of the participants not receive the program or participate in the program? • Did some comparison group members participate in the program? • Was the lack of participation in the study expected or did an unusual event occur? • If the program were implemented again, would the same kind of noncompliance arise? • Would it be possible and feasible to improve compliance when implementing the program in the future?
Counterfactual. The program being studied is similar to its comparison.	*Math Rocks* was originally sold as a new version of *Math Is Awesome* until the company decided that enough had changed to rebrand it as a new product.	• Is the program that is being used with students in the comparison group similar to the one being studied? • Is there a basis for expecting a difference in performance between the two groups?
Outcomes. The outcome measures do not consistently measure what they intend to measure.	*Math Rocks* focuses on mathematical concepts that are not part of the required curriculum in the state and are therefore not covered by the state assessment.	• Was an appropriate assessment used to measure outcomes? • Did the assessment measure skills that the program was supposed to improve? • Did the assessment accurately and consistently measure student performance? • Was the assessment measured and collected in the same way across groups? • Was the assessment created by the program developer or researcher? • Was the outcome assessment a subjective measure?

Issue	Example	Questions to Ask Yourself
Power. The ability of an analysis to distinguish an effect from pure luck can be limited by the size of its sample and how the sample was created.	Ten percent of students moved out of district during the year, and a flu epidemic caused 20 percent of students to be absent on the day of the test.	• Were some of the initially assigned study participants excluded from the analysis because of missing data or lack of consent? • Did the analysis focus on effects on subgroups of people or sites with certain characteristics? • Were the program and comparison groups very different in size? • Was the program assigned to schools or classrooms, rather than to individual students?
Baseline equivalence. Differences between the groups prior to the study may alter the findings.	In the year prior to the study, students in the *Math Rocks* group scored 5 points lower on the end of year assessment than students in the *Math is Awesome* group.	• Were the program and comparison groups formed through a random process? • Were characteristics prior to the study reported for students in the analysis? • Were the groups of students in the analysis similar prior to the study? • Did the analysis include a statistical adjustment to account for any differences in characteristics between the groups prior to the study?

What do I make of a *no effect* finding?

Seeing no issues such as those described in Table 1, a reader might conclude that there really is no effect. That means that there isn't strong evidence that any observed difference between the groups was due to the program rather than chance.

Does this mean that the program does not work or the students did not learn? Not necessarily. Instead, it means that the program may work just as well as but no better (or worse) than the program to which it was compared.

Does it mean that implementing the new program is a bad idea? Answering that question requires placing in context the finding that there was no effect on outcomes. Specifically, considering the size and implications of the observed effect may be valuable.

The observed difference may be substantively important, even if it does not meet the statistical criteria of an effect. For example, a small study may only be able to statistically detect a 20-point gain on an achievement test, although a gain of 10 points (while not statistically distinguishable from zero) may be meaningful.

Additionally, the observed difference for the full sample may mask variation in effects. For example, although a gain in achievement test score caused by the program may not attain the statistical definition of an effect overall, the program may be more effective for some types of sample members.

Some effectiveness studies provide information on teacher or student satisfaction, ease of use, supports for using the program, and the cost of implementing the program (including training and materials). Each of these factors could make a program more desirable, apart from observed

differences in student outcomes. For example, if a district is choosing between two programs with statistically indistinguishable achievement results, the district may choose a less costly program.

It can also be useful to look at the fuller body of evidence about a program from a synthesis or *systematic review*. In education, the What Works Clearinghouse (WWC) reviews effectiveness studies on education interventions, identifies the high-quality studies, and summarizes their evidence (http://ies.ed.gov/ncee/wwc). The WWC uses specific terminology to describe a lack of conclusive evidence, stating that "there are no discernible effects" for an intervention when none of the studies that examine it show statistically significant or substantively important effects.

Syntheses of studies of the same program may overcome design and resource limitations faced by individual studies. For example, multiple small studies of the same program may each have found no effect, but have results of similar direction and size across a variety of contexts. Consistency like this adds weight to arguments that a program has an effect, even if no single study has identified a statistically significant effect.

Summary

A finding of no effect is a statistical statement that an observed effect cannot be distinguished from a difference that would appear by chance. It says that a program may be just as effective as the program or programs it was being compared to, in the same way that aspirin is about as effective on average for treating headaches as acetaminophen or ibuprofen, although some individuals prefer one to another. This finding is information about the effectiveness of the program and not a conclusion about whether to implement it. A research synthesis may provide more comprehensive evidence about the effectiveness of a program than an individual study may.

Many factors determine a study's ability to detect an effect, including how well the program was implemented, the alternative to which it was compared, the similarity of the groups prior to the study, and the use of valid and reliable outcomes. Thinking about reasons for a finding of no effect may help inform improvements to the program and hypotheses for future research. A finding of no effect should be interpreted carefully with the help of information about program implementation and what the program was compared with, and it should inform decisions as one factor within a broader context, including teacher and student satisfaction, ease of use, supports for program use, and cost.

References

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., and Campuzano, L. (2007). *Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort* (NCEE 2007-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., and Doolittle, F. (2011). Middle School Mathematics *Professional Development Impact Study: Findings After the Second Year of Implementation* (NCEE 2011-4024). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., and Pistorino, C.. (2005). *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program: Final Report*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Somers, M.-A., Corrin, W., Sepanik, S., Salinger T., Levin, J., and Zmach, C. (2010). The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers (NCEE 2010-4021). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.