A Framework for Learning From Null Results

Robin T. Jacob[1], Fred Doolittle[2], James Kemple[3], and Marie-Andrée Somers[2]

A substantial number of randomized trials of educational interventions that have been conducted over the past two decades have produced null results, with either no impact or an unreliable estimate of impact on student achievement or other outcomes of interest. The investment of time and money spent implementing such trials warrants more useful information than simply "this didn't work." In this article, we propose a framework for defining null results and interpreting them and then propose a method for systematically examining a set of potential reasons for a study's findings. The article builds on prior work on the topic and synthesizes it into a common framework designed to help the field improve both the design and interpretation of randomized trials.

Keywords: evaluation; experimental design; experimental research; planning; program evaluation; research utilization

Over the past two decades, a growing portion of the education research community has dedicated itself to conducting randomized trials and other quasi-experimental evaluations of educational interventions with the goal of rigorously determining which interventions have a causal impact on outcomes for children. Of these, a substantial number have reported what some consider to be "null" results with either no impact or an unreliable estimate of impact on student achievement or other outcomes relevant to the intervention strategy. Such findings have been most common in large-scale effectiveness trials that test educational interventions in "conditions of normal educational practice" and against "business as usual" services, such as those typically sponsored by the Institute for Education Sciences (IES) National Center for Educational Evaluation (NCEE) and scale-up studies funded by IES's National Center for Educational Research (NCER). These two offices (NCEE and NCER) commissioned or funded a number of evaluations using randomized control trials between IES's inception in 2002 and 2015. A review of those studies by the Coalition for Evidence-Based Policy (2013) found that of the 77 studies implemented without any major study limitations, only 7 (9%) produced positive effects; the remaining 91% found weak or null effects. Whereas many of these studies were commissioned because there was already some reason to question whether the programs or policies were in fact reaching their intended goals, the percentage of studies yielding null results is nonetheless surprising. Education does not appear to be distinct in this regard. In medicine, for example, 50% to 80% of Stage II trials that find positive results are not able to replicate these results in larger randomized trials (Chan et al., 2008; Ioannidis, 2005; Zia et al., 2005).

Smaller scale efficacy trials in education also show nontrivial, albeit lower rates of null results. A review of studies from the What Works Clearinghouse (WWC), for example, indicates that around half of the randomized control trials that met the WWC standards and targeted mathematics or literary outcomes found null results. Compared to the NCEE and NCER studies described previously, studies reviewed by the WWC are more likely to yield positive results for two reasons. First, although there is no "file drawer problem" in commissioned studies—they are published regardless of their findings—the range of studies cataloged in the WWC likely suffers from a substantial file drawer problem. This is both because studies of interventions that do not work are less likely to get published and program developers are more likely to try to publish evaluations that show that their program is effective (Franco et al., 2014; Rosenthal, 1979). Program developers and researchers are also more likely to invest the time and money necessary to rigorously evaluate interventions that already have some evidence of effectiveness. This is true, for example, with respect to IES "Goal 3" studies, which are required to demonstrate some evidence of effectiveness before the studies are funded. This makes them

[1] University of Michigan, Ann Arbor, MI
[2] MDRC, New York, NY
[3] New York University, New York, NY

Educational Researcher, Vol. 48 No. 9, pp. 580–589
DOI: 10.3102/0013189X19891955
Article reuse guidelines: sagepub.com/journals-permissions
© 2019 AERA. http://er.aera.net

somewhat more likely than a typical intervention to produce positive impacts.

Given the investment of both time and money spent implementing rigorous evaluations of educational interventions, some have raised concern about the prevalence of these null results. This article argues that these concerns are misguided. When designed and interpreted appropriately, null results have the potential to yield valuable information beyond simply "this didn't work."

In the spring of 2015, a small conference was convened to discuss the prevalence of null results in education research, explore potential explanations for these null findings, and discuss ways to make studies more useful even when they yield null results.[1] Participants in the conference had studied a wide range of interventions using randomized designs and had all found null or mostly null results. A number of key ideas emerged from that meeting, the most important of which was that to gain purchase on understanding null findings, we need a common framework for assessing and understanding them. Establishing such a framework is the purpose of this article.

At the same time, this article argues that null findings can be as useful as studies that find positive impacts. Researchers have typically planned random assignment studies with the goal of figuring out "what works." If instead we planned studies with a goal of learning "how to make things work better," the results of any study, positive, negative, or null, could provide useful insights.

We begin by discussing what constitutes a null result, suggesting that results can actually fall into a number of different categories: positive, negative, null, or not enough statistical power to make a determination. We then offer a structure for interpreting null findings and exploring potential reasons for them. By systematically examining findings, they can be understood more clearly and in ways that can provide useful insights into methods for designing more effective educational interventions and stronger evaluations of those interventions.

This topic is not new. Researchers have approached the problem of null results from a variety of perspectives over the past decade, noting a number of potential contributors (see e.g., Landis et al., 2014; Lemons et al., 2014; Shager et al., 2013). However, the writing and research that has been done to date has tended to focus on discrete aspects of the problem (e.g., lack of statistical power, poor measurement of outcomes, implementation problems). This article is an attempt to build on this prior work and synthesize it into a common framework that can help the field improve both the design and interpretation of randomized trials.

As already noted, this article takes the perspective that null results are not undesirable. In some cases, they may represent a valid and reliable assessment that a program or policy simply did not work. Such information should be factored into decisions about whether to expand the intervention or perhaps even discontinue it. This may be particularly important when an evaluation finds that a program that is being widely used is not effective. In other cases, null effects may be an artifact of limitations in the study design. Here, a clear understanding of those limitations could be used to inform the design of future studies. In still other cases, null results may exist among a range of findings that vary across outcome measures, implementation conditions, or subgroups of participants. In all of these cases, it is important to complement the evidence with other information about the intervention, including the underlying theory of action, quality of implementation, cost, and context in which it was evaluated. With this in mind, Table 1 provides an overview of the factors that we propose researchers consider first, as they design educational evaluations and then, to revisit after the completion of a study. We refer back to Table 1 throughout the article.

Most examples in this article focus on academic achievement as measured by standardized test scores as the outcome of interest. We do this for two reasons. First, academic achievement as measured by standardized test scores is frequently the primary target of educational interventions. Second, studies with academic achievement as the outcome measure are more likely to yield null findings. Standardized test scores are, of course, not the only important outcome of interest in education, but we believe that many of the points raised here apply to a wide variety of interventions with a range of potential outcomes.

Defining a Null Result

Many evaluations of educational interventions report null findings, but such findings can and are interpreted in a variety of ways. What constitutes a null finding? Is a null finding one in which (a) the estimated impact was a precisely estimated zero, (b) the estimated impact was nonzero but not statistically significant, (c) the estimated impact was statistically significant but too small to be substantively meaningful, (d) the estimated impact was positive and statistically significant but deemed not to be worth the monetary or financial investment required to implement it, or (e) there were a mix of positive or negative and statistically significant and nonsignificant estimates across a range of important outcomes? Being clear about what we mean is an important first step in providing a framework with which to interpret findings.

In establishing a framework for assessing null results, we define a null result as one in which the estimated impact is both small and has a tight confidence interval. By small, we refer not necessarily to the magnitude of the numerical point estimate but to its practical significance. In many instances, impacts that are numerically small may still be substantively meaningful. As an example, Hill et al. (2008) provided a useful overview of the ways in which researchers can assess the substantive significance of impacts on student achievement as measured by standardized test scores.

In addition, to be considered null, impacts should be estimated with enough precision so that both the upper and lower bounds of the confidence interval are small. More generally, confidence intervals are a useful tool for interpreting findings because they contain information not only about the precision of the estimate but also about the direction and magnitude of the impacts (Aberson, 2002; Hubbard & Armstrong, 1997). If the entire span of a confidence interval contains effects that have little substantive significance, then we can more confidently conclude that the results are truly null. We argue, as have others,

Table 1
Factors to Consider in Designing Studies and Reporting Findings

	Factors	When Designing a Study…	When Reporting the Results of a Study…
Factors to consider when interpreting null results	Expected size of the estimated impact	• What impacts would be expected given typical student growth on the outcome of interest over the period of time under study? • What impacts would be expected given the intensity of the intervention and what is already known in the literature about similar interventions? • How distal is the key outcome to the intervention? How big would proximal impacts need to be to achieve impacts on distal outcomes?	• Are the results null? Is the estimated impact substantively small and measured with enough precision so that the upper and lower bounds of the confidence interval are both small? • Are the results surprising in light of typical student growth, the intensity, or distal nature of the outcome?
	Cost	• What is the intervention’s cost? Is there a minimum detectable effect per dollar that could be expected? • Can data about the cost of the intervention be collected as part of the study?	• Is the intervention substantially less expensive than the alternative to which it was being compared? • Is the cost of the intervention so low that any impact makes the investment worthwhile?
	Prior research	• Can the study be designed in a way that will make it easier to aggregate the findings with other existing studies?	• Is the null finding consistent with prior research on the intervention or similar interventions? If not, why not? • Can findings be aggregated across studies to yield more useful information?
Potential reasons for null results	Methodological factors	• Will there be adequate statistical power to detect the expected effects, considering what other large randomized trials of similar interventions have found? • Is there adequate power to detect impacts on potential subgroups of interest? • Are outcome measures appropriate to the intervention, policy relevant, measured at the appropriate time, and sensitive enough to detect effects? • What is the nature of the “business as usual” control condition? Can sites be selected to ensure a strong service contrast? Do the conditions being compared represent substantially different options for using scarce resources? • Are there factors that may influence the nature of the control condition over the course of the study, including the study itself? How can these factors be minimized? • Does the study design take contextual factors such as district churn/teacher attrition into account? • Is the study designed so that learning from real-world heterogeneity/context effects can take place?	• Were the two groups equivalent at baseline? • What is the minimum detectable effect size? Is it small enough to detect meaningful impacts? • What was the actual differential in treatment between the program and control? Was the service contrast strong? Did the counterfactual condition change over time?
	Program implementation	• What are the essential elements of the program? How will these elements be measured? • What factors need to be in place to ensure high-quality implementation? • What are the contextual factors that could impede implementation? Can sites be chosen to minimize these factors?	• Were the essential elements of the program delivered consistently to program participants? • Did the intervention reach its intended intensity? • Did the target group actually take up the services offered? • If not, why not? Was the program organized to ensure high-quality implementation? • What contextual factors were a barrier to implementation?
	Theory of change	• Does the intervention have a well-articulated theory of change? • How will the key elements of this theory of change be measured? • Does the theory of change fit the context in which it will be implemented? Can sites be chosen to ensure a better fit? • Does the theory of change suggest that the program may be more effective for some participants or in some contexts but not others? If so, will there be sufficient statistical power to ensure that that subgroup differences can be detected?	• Was the theory of change well defined? • Did the theory of change fit the context? • Did the program work for some participants or in some contexts but not others? • Do the findings suggest ways that the theory of change should be changed or modified?

that an exclusive focus on the statistical significance of effects without considering the substantive magnitude of those effects is misplaced (e.g., Wasserstein & Lazar, 2016). Studies conducted with large sample sizes can yield statistically significant results that have little substantive meaning. Small sample sizes can result in large point estimates but with confidence intervals large enough to raise doubts about the true magnitude of the findings. In this case, as we discuss in the following, the problem may be with the design of the study and not the intervention itself.

When multiple outcomes and subgroup impacts are measured but only some are positive and substantively and statistically significant, an adjustment that accounts for multiple hypothesis testing should be considered to ensure that the importance of one statistically significant finding among many is not overstated (e.g., Schochet, 2008b). Adjusting for multiple hypothesis testing will inevitably involve a tradeoff regarding statistical power, which will need to be taken into account when determining how best to present findings. Regardless, when results are mixed, researchers should be clear about the outcomes the intervention is likely to impact and for whom the intervention is most effective. A reading intervention that has positive impacts on decoding skills but not reading comprehension skills is effective at improving decoding skills but not an effective reading intervention overall; an intervention that has no main impacts but does have impacts for the boys in the sample is an effective intervention for boys but not for all students (Nosek et al., 2015).

This discussion suggests that in reality, evaluations can yield at least five different types of results: precisely estimated positive results, precisely estimated negative results, null results, imprecise estimates of program impact, or mixed results, in which findings are positive (or negative) for some outcomes or subgroups and null for others. To be considered null, findings should be both substantively small and measured with precision. In these instances, researchers can ask several questions about the findings that can help shed light on the most appropriate interpretation of the results.

Interpreting Null Findings

There are several factors to consider when interpreting null findings. First, researchers should explore whether the initial expectation that the intervention would find positive results was realistic. Although many program developers are optimistic about the potential for their intervention to produce positive outcomes, it is worth considering whether those expectations were realistic given what was known about the intensity of the intervention, how much children typically grow over the period of time under study, and how distal the key outcomes were to the intervention. Second, taking the costs of the intervention into consideration can have a substantial impact on the way null results are interpreted. Inexpensive interventions with small or even zero impact may be noteworthy simply because they are less expensive than the alternative. Finally, the finding from the study can be compared to other studies of the same or similar interventions. If the null result is consistent with prior research, then the interpretation might be different than if the null finding is anomalous.

The Expected Size of the Impact

As shown in the second column of Table 1, there are a number of factors to consider when determining a realistic assessment of the potential impact of an intervention, including what is known about typical student growth over the length of the intervention, the intervention's intensity, and the proximity of desired outcomes.

Typical student growth. One consideration is how big the impact might be expected to be given typical growth in the outcome of interest among the population of interest over the period of time the intervention was being implemented. Consider summer programming for upper elementary students as one example. How large would the expected impact of a 6-week summer program for fourth and fifth graders on standardized test scores be given that students in these grades typically only gain 0.30 to 0.40 *SD* over the course of 12 months (Hill et al., 2008)? If students were gaining as much as they do during a typical year, at best, we might only expect them to gain 0.05 *SD* over the 6-week program (0.40/12 = 0.03 *SD* of growth per month × 1.5 months = 0.05). Few studies are powered to detect impacts this small. Yet, it seems unlikely that a 6-week summer program for students in this age group could produce impacts that were substantially larger. Even if the program were twice as effective as the instruction received during a typical school year, impacts would only be around 0.10 *SD*.

As noted in Table 1, when designing studies of this kind, researchers might need to power them differently or measure outcomes in a different way. When interpreting findings, researchers and policymakers might consider whether the results were actually smaller than would have been expected given typical growth for the targeted grades, targeted domain(s), and particular subgroups of children that were the focus of the intervention.

Intervention intensity. Another consideration is the intervention's intensity. Lighter touch interventions are likely to produce smaller impacts than more intensive interventions. Comparing impacts to other programs can help establish what might be reasonable to expect. As an example, a meta-analysis of the most effective one-on-one tutoring programs for struggling readers found average impacts of around 0.60 *SD* for programs that provide intensive tutoring for 30 minutes a day 5 days a week for between 12 and 20 weeks, which is equivalent to 40 hours of additional instruction per child (Slavin et al. 2011). A program like Reading First, which provided money to schools to improve reading instruction, added around 10 additional minutes of classroom-based reading instruction per day for approximately 30 weeks of the school year (15 additional hours in total; Gamse et al., 2007). If one assumes that classroom instruction is equally as effective as one-on-one tutoring, at most, Reading First might have been expected to have impacts on reading instruction of around 0.22 *SD*. If one assumes classroom instruction is somewhat less effective than one-on-one tutoring, then the expected impact would be even smaller (perhaps around 0.10 *SD*). Again, as shown in Table 1, researchers might need to use more sensitive outcome measures or power studies differently to assess the impact of lighter touch interventions, and the intensity of the intervention should also be examined when interpreting studies with null results.

Proximity of the outcome to the intervention. Interventions that are designed to have an impact on distal outcomes likely need to produce large impacts on more proximal outcomes if they are to have a substantial impact on the outcome of interest. Teacher professional development provides a useful illustration. An article by Blazar (2015) suggested that a 1 *SD* increase in teachers' ambitious mathematics practice is associated with around a 0.10 *SD* increase in achievement. Most studies of professional development programs found relatively modest impacts on instructional practice and teacher knowledge, in the range of 0.10 to 0.60 *SD* (e.g., Garet et al., 2011; Gersten et al., 2010; Jacob et al., 2017; Schneider & Meyer, 2012). This would mean that at most, such interventions might be expected to have impacts on achievement of no more than 0.01 to 0.06 *SD*.

Likewise, a meta-analysis by Jacob and Parkinson (2015) found that after controlling for background characteristics and IQ, the largest association between executive function and achievement was around 0.15. The most effective school-based interventions designed to impact executive function have only had impacts on measures of executive function equal to around half a standard deviation (e.g., Raver et al., 2011). This means that under the best case scenario (e.g., a treatment impact of 0.50 *SD* on measures of executive function and a true association between executive function and achievement equal to 0.15), interventions designed to improve executive function would only have the potential to increase future achievement by less than a tenth of a standard deviation (half of 0.15).

As Table 1 notes, these factors should be taken into account when designing a study, and studies should be powered appropriately or outcomes measured differently (perhaps using more sensitive measures or measuring impacts over a longer period of time). Such factors should also be considered when interpreting findings from a study that has a small and precisely estimated null impact. Short-term studies of interventions that are designed to impact student achievement indirectly are likely to yield only small impacts on standardized test scores.

Considering Cost

Researchers often fail to take an intervention's costs into account when interpreting a study's findings. Even in instances where the impact of the intervention is a precisely estimated zero, if it is substantially less costly than the business as usual condition to which it is being compared, a zero impact might be considered a success. Consider a study that compared two different approaches to helping struggling readers: small group intervention support and computer-based tutoring. If the evaluation were to find no impact of the computer-based tutoring, which has a very low cost per student, compared with small group intervention support, which is quite costly, one might conclude that the computer-based tutoring is a success because it achieves the same results at a lower cost.

Similarly, a very low-cost intervention with small impacts might be considered successful despite small impacts precisely because the cost is low. For example, a study by Glewwe et al. (2014) showed that providing eye exams and free eyeglasses to students who need them (at a cost of approximately $7/student) had a positive impact on test scores of between 0.07 and 0.16 *SD*. This is somewhat smaller than the impact of the Tennessee STAR class size experiment (which found impacts in the range of 0.20–0.25 *SD*) yet substantially less expensive (Krueger, 1999). Krueger (1999) estimated that the cost (in 1996 dollars) of reducing class size by one-third would be about $2,151 per student per year. This idea is closely related to the idea of equivalence testing, an alternative to traditional hypothesis testing. An equivalence test establishes a confidence interval that statistically rejects the presence of effects large enough to be considered worthwhile (see Lakens, 2017, for a discussion).

"Academic mindset" training provides a similar example. Mindset training targets students' core beliefs about school and learning and has been shown to have a small positive impact on students' GPA and successful course completion (Paunesku et al., 2015). Because the intervention is delivered via a computerized program, the developers noted that the intervention could be scaled to virtually unlimited numbers of students at a very low marginal cost. Thus, even though the impact of the intervention is relatively small (an increase in the rate of satisfactory performance in core courses of 6.4 percentage points), scaling it could result in hundreds of thousands more successful course completions among low-performing students. The Center for Benefit-Cost Studies of Education at Teachers College provides useful resources for researchers wishing to systematically estimate an intervention's total cost (http://cbcse.org). As noted, in Table 1, an explicit focus on cost should be incorporated into study designs so that the appropriate information about the costs of the intervention is collected and should be discussed when reporting the results of the study.

Placing Findings in Context

Finally, as noted in Table 1, the results of a single study should never be considered in isolation. Today, there are often many evaluations of the same intervention or the same type of intervention that can be used to place the findings of one study in context. For example, is a null finding consistent with prior research on the intervention, or have prior evaluations found positive impacts? If it is unique, what factors might have contributed to the null result in this instance? Can the results of multiple studies be aggregated so that a more robust estimate of impacts can be determined or help identify contexts or subgroups for which the intervention is most effective?

Exploring the Potential Causes of the Findings

Even after taking these factors into consideration, however, researchers, evaluators, and program developers may conclude that the impacts of a program were either small or truly null. In such cases, researchers must explore the potential causes for these findings.

Methodological Factors

The first question is whether the conditions for a strong impact evaluation existed or whether the findings might be explained by weaknesses in the initial study design or a poorly implemented study design. Many of these considerations, including whether the sample size was adequate to detect effects, whether there was differential attrition between the treatment and control group, and whether the two groups were comparable at the start of the intervention, have been covered extensively in the literature. We review a few of the most common methodological factors here.

Statistical power. Studies are often designed without sufficient statistical power to detect program impacts even if they exist. Spybrook and Raudenbush (2009) examined power analyses for the first wave of group-randomized trials funded by the IES. They found that studies that were funded earliest (between 2002 and 2004) had minimal detectable effect sizes between 0.40 and 0.90, which means that they likely were not sufficiently powered to detect meaningful effects (Lipsey & Wilson, 1993; Schochet, 2008a; Spybrook & Raudenbush, 2009). The precision of the studies improved over time such that studies funded between 2006 and 2008 had minimum detectable effects ranging from about 0.18 to 0.40; however, as noted previously, given the nature and intensity of interventions, these may still not have provided enough statistical power to detect effects that could be considered meaningful. Spybrook (2014) also showed that studies typically do not have enough precision to detect differences in intervention effects across key subgroups of interest, such as rural and urban schools or high- and low-poverty districts, meaning that it will be difficult to identify if programs are effective for some subgroups even if they are not effective overall.

Typically, evaluations are powered to detect effects in the range of 0.20 to 1.0 *SD* (Spybrook & Raudenbush, 2009). Researchers often use estimates of program effectiveness from small, experimental, and nonexperimental evaluations of the intervention of interest to estimate the likely impact of the program. However, as has been shown in education and other fields, sample size has a strong negative correlation with effect size (Slavin & Smith, 2009). As will be discussed in more detail later, this may be in part because such interventions are often implemented under ideal conditions, with close oversight from the program developers, or because small studies with positive impacts are more likely to get published. These studies probably overestimate the size of the impact that is likely to be found in larger, more rigorous studies. Investigators might do better to use larger randomized trials of similar programs as a benchmark for establishing a minimum detectable effect rather than using smaller, less rigorous evaluations of their own programs.

As noted earlier, and as shown in Table 1, a broader range of factors needs to be considered to help identify a credible minimum detectable effect size (MDES) around which to design a study. Factors to consider include children's developmental trajectories in different domains and different contexts as well as the nature and intensity of the treatment. This approach not only provides a sense of what is reasonable to expect an intervention to achieve, it also requires that we think carefully about the counterfactual condition to which the intervention is being compared (e.g., How much would one typically expect a given population to grow over this period of time? What other services is the control condition receiving in the absence of treatment?) and the nature of the treatment itself (e.g., How is the intervention going to achieve the desired results? How realistic are the assumptions behind it?). All of these considerations should lead to more thoughtful designs, grounded in real-life assumptions.

Outcome measures. A substantial amount of attention has also been given to the importance of measurement of outcomes in randomized trials. Both measurement error and missing data can increase the sample size required to achieve a minimum level of precision, thus reducing the likelihood of finding effects (R. B. Olson et al., 2011). The timing of assessments can also impact the degree to which impacts can be detected; if pretest data are collected after the start of the intervention or posttests are administered before the intervention has concluded, estimates of program impact may be smaller than they otherwise might have been (e.g., Schochet, 2010). Similarly, some program benefits may not be realized until many years after the intervention has been completed, and assessing only the short-term outcomes may mask the longer-term benefits that may accrue to participants. This has been best illustrated in the case of high-quality preschool, where long-term follow-up studies have revealed impacts on a variety of outcomes, including a reduction in grade retention, special education placements, delinquency, and incarceration rates many years after the completion of the program (e.g., Barnett & Hustedt, 2005).

Finally, the degree of alignment between the treatment and the outcome measure will influence the size of the impact estimate, with highly aligned measures more likely to yield larger impact estimates than those that are less closely aligned and proximal measures more likely to yield positive effects than more distal measures (e.g., Hill et al., 2008; Marzano, 2014; R. B. Olson et al., 2011; Slavin & Madden, 2011). Slavin and Madden (2011) showed that among WWC studies of beginning reading programs, the outcome measures that were closely aligned with the intervention yielded weighted effect sizes of 0.51 *SD*, whereas the more general outcome measures had average effect sizes equal to 0.06. In math, the average effect sizes were 0.45 and –0.03, respectively. Considering whether the right outcomes are being measured and subsequently whether the measures that were used have a bearing on the results of a study can help reduce the likelihood of finding null results and shed light on the potential reasons for them (see Table 1).

Treatment-control contrast. In addition to these well-documented concerns about study design, there has also been a recognition that the counterfactual to which the intervention is being compared plays a critical role in the magnitude of effects. In their seminal work, Shadish et al. (2001) identified the counterfactual condition as one of the key potential threats to the validity of experimental and quasi-experimental designs. In recent years, there have been a number of examples that have underscored the importance of the counterfactual condition. For example, Shager at al.'s (2013) research on variation in the results of Head Start found that studies that had an active control group, in which control children were enrolled in other center-based programs, had much smaller effect sizes than studies in which the control group received no other early childhood education. They noted that according to Cook (2006), almost 70% of 4-year-olds attend some form of early childhood education, thus increasing the likelihood that program impacts will be diluted by a counterfactual condition in which the control group children are receiving services.

In a similar vein, D. Olson (2004) suggested that the counterfactual is likely to both change over time or change in response to the treatment, making it difficult to identify the factors that "uniquely define the treatment" itself. This phenomenon was demonstrated empirically by Lemons et al. (2014), who explored data from five randomized control trials of the Kindergarten Peer-Assisted Learning Strategies program, a supplemental, peer-mediated reading program. The study showed that over the 8 years in which the studies took place, there was a dramatic increase in the performance of control students over time, which substantially reduced the observed impacts of the program, despite the fact that students in the treatment group showed substantial gains in their literacy skills. The authors suggested that the increased performance of the control group might be attributed to a changing national and district policy landscape with increased emphasis on literacy instruction in kindergarten.

The presence of the intervention can also cause schools to redirect their resources in ways that change the counterfactual condition. A recent evaluation of the Reading Partners volunteer tutoring program for struggling readers, in which students were randomly assigned within schools to receive the program or to a business as usual control group, found that students in the control group were more likely to receive small group intervention services than students in the treatment group, suggesting that schools were shuttling their additional resources to the control group. This resulted in a situation in which 65% of the control group was also receiving some type of supplemental reading instruction, likely diluting the estimated impact of the intervention (Jacob et al., 2016).

In the case of most educational interventions, a "no services" control group is typically not a realistic option, nor does it provide the right comparison to address policy-relevant questions. At one point in time, a no services control group might have been possible for young children, but the prevalence of preschools has proliferated over the past several decades, and today, most children attend some sort of early childhood program. A no services control group was never a realistic option in the K–12 environment. A literacy program will always be compared to the existing literacy program that students would otherwise have received, for example. Therefore, most evaluations of educational interventions are tests of differential impact. In many cases, such a comparison addresses the policy-relevant question—how does the new intervention or approach compare to what is typically done? However, if we expect to see differences in outcomes, then the comparisons must be between substantively different ways of doing things.

For these reasons, a careful assessment of the counterfactual is needed prior to the implementation of any evaluation. What is the current business as usual condition? What will the intervention that is being tested add above and beyond the current condition? Do the conditions being compared represent options between important alternative choices for investing scarce resources? Will the implementation of the intervention or the study change the counterfactual condition? Are there other factors at play that could change the nature of the counterfactual over time? As highlighted in Table 1, to answer questions about the extent of the actual service contrast requires program developers and evaluators to identify those aspects of the intervention that are most likely to drive impacts and collect detailed implementation data from both the treatment and control groups throughout the intervention period.

Program Implementation

Studies in which the impact of the intervention was truly null, the study design and methodological conditions for a strong impact evaluation were in place, and there was a sufficient contrast between the services or intervention received by the treatment and control group raise additional questions. The first is whether the program model was implemented with fidelity and the intended dosage. Table 1 indicates that to assess this requires that the essential elements of the program be identified and measured. What are the core elements of the program, and were those elements delivered consistently to program participants? And if not, what were the potential explanations? Was the program organized to ensure high-quality implementation? For example, were structures in place to ensure redundancies? Was there sufficient monitoring of those tasked with the implementation? Were there contextual factors that served as a barrier to implementation? Or was the program simply too complicated to implement effectively in real-life conditions?

Answering these questions requires careful measurement of the key aspects of implementation, including the amount of exposure provided, the quality of delivery, and the ability of those on the frontline to adapt to changing circumstances and contexts. Yet, these aspects are often not well thought out in the design of the study, and as a result, findings are difficult to interpret. Furthermore, ideally, we would want to know which of these aspects of implementation matter more or less to the success of a particular program, but these factors are often not considered in the initial design of the study.

These factors are particularly important when evaluating programs that are being implemented at scale because to be successful, such programs need robust implementation plans that do not require unusual supports. Documenting the number and types of implementation barriers that treatment sites face and assessing contexts where implementation is more or less challenging could yield useful information for future implementation even in the context of null overall findings. For example, an evaluation of the ANet program divided matched pairs of schools into three "readiness" groups (top, middle, and bottom) based on ratings by program staff regarding their readiness to engage in instructional data use and found that schools in the lowest readiness category had negative impacts on student achievement, while those in the highest category had positive impacts, even though the overall results of the evaluation were null (West et al., 2016). Thus, future implementations of the program could screen for readiness factors as a prerequisite for program participation.

Contextual factors as a barrier to implementation. Although there is growing recognition in the field regarding the importance of implementation as a key contributor to program effectiveness (e.g., Goodson, 2015), the role that contextual factors play in helping or hindering implementation is often overlooked. By contextual factors, we mean elements in the system that impact the intervention or the interaction of individuals within the system to the detriment of the program. Changes in district leadership, shifting district priorities or incentive systems, or lack of principal support for an intervention can all impact how teachers implement an intervention or respond to a reform model. Lee et al. (2013), in their study of the impact of No Child Left Behind school interventions in New York, noted that overlooking contextual factors, such as the social and racial composition of the school or a district's internal capacity for change, can prove "fatal" in the implementation of any reform effort. These contextual factors were a key factor that distinguished between successful and unsuccessful efforts at reform in their study.

Elmore (1996) similarly argued that it is school organizational structures and incentive systems that make it difficult to change the core practices of teaching. This same sentiment was echoed by D. Olson (2004), who suggested that a closer examination of "the school's place in the institutional structures of a bureaucratic society and the categories and rules, knowledge and procedures, that are required for successfully participating in it" (p. 25) would yield greater insights into the reasons for the numerous null findings in educational research. And Wilson (2013) noted that although existing research on effective professional development identified factors that led to effective professional development, the complexity of the U.S. educational system often thwarts efforts to support teachers, and as a result, interventions that take a more systemic approach are more likely to be effective.

There may also be an interaction between the context and the evaluation itself. For example, in school-level randomized controlled trials, teachers may have a sense that the intervention or program being studied is just a special project that will be over soon, so they need not take it seriously or invest in it fully.

Theory of Change

Situations where the results can be considered truly null, the study design was strong, and the program was implemented with a high degree of fidelity to the program model suggest that there may be flaws in a program's theory of change that led to disappointing results. This potential explanation may be the most difficult to embrace because it often challenges a priori assumptions and long-held beliefs. Yet, it may also be the most useful for expanding knowledge about the most effective ways to intervene in children's lives.

In some instances, the theory of change may simply be wrong—the fundamental understanding of the factors that contribute to teachers' or children's learning or development may be flawed. However, other scenarios are also likely.

For example, the theory of change may be generally right but only under ideal conditions (e.g., only with highly skilled staff, only with a substantial amount of oversight and training, etc.). Slavin and Smith (2009) cited this as one of the reasons that small randomized trials (e.g., with sample sizes with fewer than 100 students) are much more likely than studies with larger sample sizes to yield large, positive, and statistically significant findings. The authors noted that the ability to closely monitor the implementation of programs that are delivered to small groups of students likely contributes to this pattern of findings. Although other factors likely play a role in this phenomenon as well, researchers have long recognized the problems of bringing effective programs to scale (e.g., Elmore, 1996).

Another possibility is that the theory of change works only in certain contexts or for certain populations. For example, recent work by Bloom and Weiland (2015) showed that past estimates of the effectiveness of Head Start programs, which focused only on overall average impacts, masked a wide range of relative program effectiveness for specific subgroups of students, particularly for English Language Learners and those beginning the program with the weakest skills. This underscores the importance of powering studies sufficiently so that subgroup impacts can be detected.

This also suggests the importance of exploring variation in impacts, with an emphasis on identifying causal contributors to the variation. This can be difficult because there are many factors that might account for variation in impacts but whose influence cannot easily be causally identified. For example, some programs may be more effective with highly skilled teachers, but teaching skill is often confounded with the treatment. Over the past few years, a number of articles have addressed potential approaches to studying such variation in impacts, and as a result, more studies are likely to explore such variation in the future (see Schochet et al., 2014, for a review). Table 1 outlines some of the ways these considerations can be factored into both the design of educational evaluations and the interpretation of findings.

Conclusions

All studies, even those with null effects, contain important information. Capitalizing on the information they contain can help guide both the design and evaluation of interventions in the future. As summarized in Table 1, study findings, particularly those with null results (i.e., studies where the point estimates on the outcomes of interest are both substantively small and have tight confidence intervals), should be interpreted with a variety of factors in mind. These include factors that may shed light on the appropriate interpretation of the findings, such as what was reasonable to expect the intervention to achieve at the outset, the intervention's costs, and what other studies of the same or similar interventions have found. They also include a consideration of the potential causes of the null findings, including (a) weaknesses in the study design, (b) whether the intervention was implemented well, (c) the context in which the intervention was implemented, and (d) the underlying theory of change behind the intervention. Considering these factors will provide a more nuanced and useful way to understand findings, particularly those with weak or null results.

At the same time, a consideration of these factors has implications for the design of future studies. The field currently designs studies expecting to find positive impacts of the program being evaluated. If instead we designed our studies in preparation for weak, null, or even negative findings—asking what we would want to know if the evaluation found that the intervention had no impact or a negative impact on the outcomes of interest—our studies, even those that yield null results, might be better situated to add useful information to the field.

This article has focused mostly on results from individual studies. Yet, individual studies are generally not the most appropriate unit for drawing summary conclusions about the effectiveness of an intervention. Meta-analysis has shown there is often substantial variability in the observed effects across multiple studies of the same or very similar interventions (e.g., Lipsey, 2009; Slavin et al., 2011). Replication is important (Makel & Plucker, 2014), and in addition to thinking more carefully about the findings of individual studies, broader conclusions about an intervention's effectiveness should be made by systematically reviewing and integrating findings based on multiple studies. As noted earlier, meta-analysis can be used to not only establish more robust estimates of program impact but also to explore variation in impacts based on many of the factors identified previously, including variation in study design, program implementation, and context.

NOTES

The research reported here was supported by the National Science Foundation through Grant No. 0918383 to Harvard University. The opinions expressed are those of the authors and do not represent views of the National Science Foundation.

[1] Conference on Null Results convened by Robin Jacob, Heather Hill, Stephanie Jones, and James Kim, May 7, 2015, Washington, DC. Supported by the National Science Foundation (DRL-0918383).

REFERENCES

Aberson, C., (2002). Interpreting null results: Improving presentation and conclusions with confidence intervals. *Journal of Articles in Support of the Null Hypothesis*, *1*(3), 36–42.

Barnett, W. S., & Hustedt, J. T. (2005). Head Start's lasting benefits. *Infants & Young Children*, *18*(1), 16–24.

Blazar, D. (2015). Grade assignments and the teacher pipeline: A low-cost lever to improve student achievement? *Educational Researcher*, *44*(1), 213–227. https://doi.org/10.3102/0013189X15580944

Bloom, H., & Weiland, C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study*. MDRC Working Paper. New York, NY.

Chan, J. K., Ueda, S. M., Sugiyama, V. E., Stave, C. D., Shin, J. Y., Monk, B. J., Sikic, B. I., Osann, K., & Kapp, D. S. (2008). Analysis of phase II studies on targeted agents and subsequent phase III trials: What are the predictors for success. *Journal of Clinical Oncology*, *26*(9), 1511–1518.

Coalition for Evidence-Based Policy. (2013). *Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects*. http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf

Cook, T. (2006). *What works in publicly funded prekindergarten education?* [Paper presentation]. Children's Achievement: What the Evidence Says About Teachers, Pre-K Programs and Economic Policies Policy Briefing. Washington, DC.

Elmore, R. F. (1996). Getting to scale with good education practice. *Harvard Educational Review*, *66*(1), 1–27.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Gamse, B., Bloom, H., Kemple, J., & Jacob, R. (2007). *Reading First impact study: Interim report*. Washington, DC: Institute for Education Sciences.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., & Warner, E. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011–4024). ERIC. http://eric.ed.gov/?id=ED519922

Gersten, R., Dimino, J., Jayanthi, M., Kim, J., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, *47*(3), 694–739.

Glewwe, P., West, K., & Lee, J. (2014). *The impact of providing vision screening and free eyeglasses on academic outcomes: Evidence from a randomized trial in Title 1 elementary schools* [Unpublished manuscript].

Goodson, B. (2015, December 1). *Evidence at the crossroads pt. 5: Improving implementation research*. http://wtgrantfoundation.org/evidence-at-the-crossroads-pt-5-improving-implementation-research

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Hubbard, R., & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports*, *80*(1), 337–338.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*(2), 218–228.

Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness*, *9*(S1), 67–92. https://doi.org/10.1080/19345747.2016.1138560.

Jacob, R., Hill, H., & Corey, D. L. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research in Educational Effectiveness*, *10*(2), 379–407.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*(4), 512–552. https://doi.org/10.3102/0034654314561338

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, *115*(2), 497–532.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Landis, R. S., James, L. R., Lance, C. E., Pierce, C. A., & Rogelberg, S. G. (2014). When is nothing something? Editorial for the null results special issue of *Journal of Business and Psychology*. *Journal of Business and Psychology*, 29(2), 163–167.

Lee, J., Shin, H., & Amo, L. (2013). Evaluating the impact of NCLB school interventions in New York State: Does one size fit all? *Education Policy Analysis Archives*, *21*(67), 1–39. http://dx.doi.org/10.14507/epaa.v21n67.2013

Lemons, C., Fuchs, D., Gilbert, J., & Fuchs, L. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242–252. https://doi.org/10.3102/0013189X14539189

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders*, *4*(2), 124–147.

Lipsey, M. W., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181–1209. http://dx.doi.org/10.1037/0003-066X.48.12.1181

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316.

Marzano, R. J. (2014). Proximal versus distal validity coefficients for teacher observational instruments. *The Teacher Educator*, *49*(2), 89–96. https://doi.org/10.1080/08878730.2014.885783

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., . . . Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Olson, D. (2004). The triumph of hope over experience in the search for "what works": A response to Slavin. *Educational Researcher*, *33*(1), 24–26.

Olson, R. B., Unlu, R., Jaciw, A. P., & Price, C. (2011). *Estimating the impacts of educational interventions using states tests or study-administered tests*. (NCEE 2012-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Paunesku, D., Walton, G., Romero, C., Smith, E., Yeager, D., & Dweck, C. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, *82*(1), 362–378.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Schneider, C. M., & Meyer, J. P. (2012). Investigating the efficacy of a professional development program in formative classroom assessment in middle school English language arts and mathematics. *Journal of Multidisciplinary Evaluation*, *8*(17), 1–24.

Schochet, P. (2008a). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87.

Schochet, P. (2008b). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. (2010). The late pretest problem in randomized control trials of education interventions. *Journal of Educational and Behavioral Statistics*, *35*(4), 379–406.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. http://ies.ed.gov/ncee/edlabs.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shager, H. M., Schindler, H. S., Magnuson, K. A., Duncan, G. J., Yoshikawa, H., & Hart, C. M. (2013). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis*, *35*(1), 76–95.

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26.

Slavin, R. E., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370–380.

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education*, *82*(3), 334–357.

Spybrook, J., & Raudenbush, S.W. (2009). An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318.

Wasserstein, R., & Lazar, N. (2016) The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

West, M. R., Morton, B. A., & Herlihy, C. M. (2016). *Achievement Network's Investing in Innovation Expansion: Impacts on educator practice and student achievement*. Cambridge, MA: Center for Education Policy Research.

Wilson, S. (2013). Professional development for science teachers. *Science*, *340*(6130), 310–313.

Zia, M. I., Siu, L. L., Pond, G. R., & Chen, E. X. (2005). Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. *Journal of Clinical Oncology*, *23*(28), 6982–6991.

AUTHORS

ROBIN T. JACOB, PhD, is a research associate professor at the Institute for Social Research at the University of Michigan, 426 Thompson St., Ann Arbor, MI 48104; *rjacob@umich.edu*. Her research focuses on the evaluation of social programs to improve the lives of young people.

FRED DOOLITTLE, PhD in economics, is a senior fellow at MDRC, 200 Vesey St., New York, NY 10281; *fred.doolittle@mdrc.org*. His research focuses on program evaluations of education instructional strategies.

JAMES KEMPLE, EdD, is executive director of the Research Alliance for New York City Schools at New York University, 285 Mercer St., 3rd Floor, New York, NY 10003; *james.kemple@nyu.edu*. Dr. Kemple's research focuses on high school reform efforts, assessing performance trends in New York City's educational landscape, and designing rigorous impact evaluations.

MARIE-ANDRÉE SOMERS, EdD, is a senior research associate at MDRC, 11965 Venice Blvd., Suite 402, Los Angeles, CA 90066; *marie-andree.somers@mdrc.org*. Her research focuses on experimental and quasi-experimental evaluations of school-based academic and social and emotional interventions for middle school students.

Manuscript received September 27, 2016
Revisions received September 12, 2017;
December 16, 2018
Accepted November 7, 2019