


Chapter 8

Toward a Science of Failure Analysis: A Narrative Review

CLAIRE ALLEN-PLATT 
CLARA-CHRISTINA GERSTNER
ROBERT BORUCH
ALAN RUBY
University of Pennsylvania

When a researcher tests an educational program, product, or policy in a randomized controlled trial and detects a significant effect on an outcome, the intervention is usually classified as something that “works.” When expected effects are not found, there is seldom an orderly and transparent analysis of plausible reasons why. Accumulating and learning from possible failure mechanisms is not standard practice in education research, and it is not common to design interventions with causes of failure in mind. This chapter develops Boruch and Ruby’s proposition that the education sciences would benefit from a systematic approach to the study of failure. We review and taxonomize recent reports of large-scale randomized controlled trials in K–12 schooling that yielded at least one null or negative major outcome, including the nature of the event and reasons (if provided) for why it occurred. Our purpose is to introduce a broad framework for thinking about educational interventions that do not produce expected effects and seed a cumulative knowledge base on when, how, and why interventions do not reach expectations. The reasons why an individual intervention fails to elicit an outcome are not straightforward, but themes emerge when researchers’ reports are synthesized.

Rhetoric on education in the United States often targets large systems and subsystems of states, districts, and schools. The scientific evidence on what works and does not work in education, however, tends to cover local evaluations of individual programs, pedagogies, instructional materials, classroom technologies, and policies

Review of Research in Education

March 2021, Vol. 45, pp. 223–252

DOI: 10.3102/0091732X20985074

Chapter reuse guidelines: sagepub.com/journals-permissions

© 2021 AERA. journals.sagepub.com/home/rre

that shape a learning environment. Interventions, attempts to improve practice or learning outcomes, are what education researchers usually test, often in randomized controlled trials (RCTs).

When a researcher tests an educational program, product, or policy in an RCT and detects a statistically significant effect on an outcome, the intervention is usually classified as something that works. It is construed as potentially worthy of replication, emulation, or scale-up. When expected effects are not found, however, the intervention is often labeled a failure or, more cautiously, as a “null finding.” Null findings in such research are often treated as an evidential dead-end. More important, there is seldom an orderly and transparent analysis of plausible reasons why the intervention did not work as anticipated. This is despite the fact that experiments are a dependable “way of establishing a cumulative tradition” of promising educational practices (Campbell & Stanley, 1963, p. 2).

This chapter develops Boruch and Ruby’s (2015) proposition that the education and social sciences would benefit from a systematic approach to the study of failure. The rationale is that we can learn from nonsuccesses in educational experiments and as a consequence improve the quality of research and the quality of interventions to which the research is directed. In what follows, we define *failure* specifically as the failure to reach a probabilistic threshold, such as .05 or less, in a formal statistical test of a null hypothesis, when an ex ante power analysis had been based on an effect size or “minimum detectable effect size” specified by the researcher and a sample size selected based on that calculation. In other words, this may be labelled as “null findings,” “no difference findings,” “no discernible effects,” “expected effects were not detected,” and so on.

The chapter reviews published reports of RCTs in which the reports’ authors (a) detected at least one null or nonsignificant major outcome of the intervention tested and (b) gave possible reasons for the failure to detect expected effects. This review is confined to large-scale RCTs in K–12 schooling from the last 10 years. The intent is to build an initial, broad framework for thinking about educational interventions that do not produce expected effects in controlled trials and to seed a cumulative knowledge base on when, how, and why interventions do not reach expectations. The chapter contributes to this volume of *Review of Research in Education* by systematically investigating an underexploited source of high-quality evidence and identifying themes that emerge through a synthesis of researchers’ accounts.

We first lay out the context by describing briefly how failure is often handled in the education sciences literature. This is in contrast to models of how the topic of failure is approached in other sciences, which are also then considered. A description of the methods used to search the literature and the resulting assembly of studies are outlined. The main sections of the chapter cover our analysis of reported reasons for the failure to detect expected effects. Implications of this work for researchers, for better ex ante design of trials and better ex post facto analysis, are covered last along with conclusions.

NULL FINDINGS IN THE EDUCATION SCIENCES

The population prevalence of null findings in education research is unknown. Estimates based on different samples vary from one third (H. C. Hill & Erickson, 2019) to approximately one half of educational experiments (Jacob et al., 2019; Kim, 2019) and 9 out of 10 follow-up effectiveness trials specifically (Coalition for Evidence-Based Policy, 2013; Kim, 2019). Null results in the social sciences may be undercounted for many reasons, including reporting and publication biases that induce researchers and editors to only publish successful (i.e., significant) statistical outcomes (e.g., Bakker et al., 2012; P. Dawson & Dawson, 2016; Franco et al., 2014; Pigott et al., 2013; Rosenthal, 1979). For example, Pigott et al. (2013) reviewed reports of educational interventions that appeared in both dissertations and published formats and found nonsignificant outcomes were 30% more likely than statistically significant outcomes to be removed from the published versions. These studies constitute partial support for Rosenthal's (1979) claim that behavioral researchers' tolerance for null results is so low as to condemn most studies with nonsignificant effects to the bottom of a file drawer.

Researchers in the applied social sciences seldom unpack why or how an intervention failed to meet expectations in a well-executed experiment, despite the commonness of RCTs with one or more null results. Boruch and Ruby (2015) and a recent special issue of *Educational Researcher* edited by Herrington and Maynard (2019) are distinct in the literature for interrogating the nature and consequence of null findings in educational experiments. Herrington and Maynard's (2019) volume asks, "What does it mean when an evaluation produces null findings? Why are null findings so prevalent? How can they be used to advance knowledge?" (p. 577), while Boruch and Ruby (2015) reflect,

How do we design randomized controlled trials *a priori* so as to better learn from the inevitable failures to meet expectations about the effectiveness of the interventions? How can we learn about plausible reasons for failure to meet expectations *post facto* in a scientific and orderly way? (p. 6)

The *Educational Researcher* issue challenges researchers to manage their expectations with respect to treatment effects—to "articulate . . . realistic assessments of the program theory being tested, the likely counterfactual condition, and the likely size of impacts" (C. J. Hill, 2019, p. 609) and "get used to the idea that the effects of educational interventions on valued outcomes are usually going to be small" (Valentine, 2019, p. 612). Contributors counsel researchers to give more practical weight to precisely estimated null or negative effects versus imprecise estimates.

This chapter addresses Boruch and Ruby's questions directly, as evinced in an assembly of published reports of RCTs, and derives from this assembly several implications for researchers that confirm and overlap with the work in Herrington and Maynard (2019), especially the framework to learn from null findings proposed by Jacob et al. (2019). Such implications include articulating an intervention's logic model; employing a small number of directly program-relevant outcome measures;

acknowledging that realistic effect sizes in school settings are often small; and collecting more data on fidelity of implementation than convention stipulates (C. J. Hill, 2019; H. C. Hill & Erickson, 2019; Jacob et al., 2019; Kim, 2019; Valentine, 2019). And we newly add to the literature the relative frequency of reasons for failed outcomes offered by researchers themselves and strategies to minimize failure potential in the design, implementation, and analysis stages of experiments.

Literature on null findings in the education sciences may be lacking because the field lacks incentives for researchers to publish and interrogate such results. Outlets for research on educational interventions, such as the What Works Clearinghouse (WWC) in the United States or the “Practices that Work” arm of the European Platform for Investing in Children, are explicit about their priority to report on programs that have some positive outcomes—programs that “work.” Schweppenstedde and Reid (2015) suggest that legislators who repeatedly underscore their use of evidence in policymaking portray evaluation studies as “a continuous success story from the stage of conception of an intervention, to its application and finally, to the impact on end users.” Reporting a nonsignificant or negative intervention outcome can be awkward for program developers, implementers, stakeholders, funders, or researchers if they are associated with a failed hypothesis or an unsuccessful investment.

The awkwardness is often more marked when the unrealized success of a substantial investment involves children (Dahlin et al., 2018). For example, the recent Gates Foundation Intensive Partnerships for Effective Teaching grants program, a multi-year, multiprogram, seven-site initiative to improve teacher quality in service of greater student achievement, resulted in almost no improved student outcomes at any site (Stecher et al., 2019). Trade publications and newspapers translated the effect estimates as “one of the biggest failures yet in K–12 philanthropy” (Hall & Callahan, 2018) and a hundred-million-dollar “bust” (Strauss, 2018). These are powerful allegations to endure without reputational and ego loss.

A general aversion to null findings might impel researchers to “define failure out of existence or define it in such a way that reduces its ostensible frequency” (Boruch & Ruby, 2015, p. 2). For example, recent controversies over how states count high school graduates (e.g., Alvarez & Marsal Holdings LLC, 2018; USDOE Office of Inspector General, 2018a, 2018b) or measure student proficiency in reading or mathematics (e.g., Achieve, 2018; Bandeira de Mello et al., 2019; U.S. Chamber of Commerce, 2007) can shape the interpretation of increases in numbers of high school completers or changes in academic skill levels of graduates. The perception of success changes even though the data are unchanged.

Changes over time in statistical and practical interpretations of failure can also be used by researchers to dodge nonsuccess. A failure to detect effects for an intervention in one decade, for instance, may be put aside when new success criteria or policy priorities emerge in another decade. To borrow an example from Maynard (2006), policies to expand the supply of childcare in the 1980s considered the primary outcome to be women’s increased participation in the workforce. In contrast, such

programs in the 1990s and 2000s primarily aimed to close school readiness gaps between socioeconomic subgroups of children and bypassed maternal employment entirely. An inverse example is present in policies to improve the quality of curriculum and assessment in U.S. K–12 schools in the 2000s, such as the Common Core State Standards (CCSS), released in 2010 (Core Standards, 2019). Curriculum and assessment products considered successful by pre-CCSS trials and outcomes (e.g., Pearson’s enVisionMATH Program, evaluated by Resendez et al., 2009) have been reevaluated under new criteria for quality and success and removed from the list of endorsed instructional materials in some states, such as Louisiana (Louisiana Believes, 2019a, 2019b).

It seems reasonable to develop more transparent approaches to understanding failure, in light of the reporting and publication bias, possible reputational harm associated with failure to detect expected results, and subjective data-analytic decisions that influence results. A shift to understanding failure could encourage transparent research and lessen the stigma associated with null or negative results, as well as bring to light accounts of unsuccessful educational experiments that are an underexploited source of evidence for practitioners and researchers.

HANDLING FAILURE AND FAILURE ANALYSIS IN OTHER SCIENCES

Other disciplines model what failure analysis as a professional habit can look like. For instance, the field of organizational behavior describes “failure learning” as a formal modality used by the transportation, medical, and manufacturing sectors to translate an adverse event (such as an accident or fatality) into an opportunity to improve quality and efficiency, transfer successful routines, and otherwise reduce the long-run frequency of adverse outcomes (Dahlin et al., 2018). Consider the following.

Formal systems of failure learning in other fields and disciplines might inform the social sciences’ examination of the topic. For example, engineering, aviation, medicine, and other fields use error typologies to standardize language around failure events and classify their severity for analysis and learning. In some literatures,

errors are incorrectly executed tasks or routines (such as . . . a nurse who gives the incorrect medication to a patient), whereas *failures* are undesired performance outcomes (. . . a patient who dies after surgery instead of leaving the hospital healthier than before entering it). (Dahlin et al., 2018, p. 254, emphasis added)

In aviation, postmortem analyses differentiate between incidents (no fatalities) and accidents (people died) to quickly communicate relevant detail (Bureau of Transportation Statistics, 2020). Likewise, educational trials might differentiate, terminologically, unrealized major outcomes (such as student-level reading achievement when the intervention is an instructional software program used by students) versus indirect or secondary outcomes (such as student-level reading achievement when the intervention is a summer professional development program for literacy teachers, or

outcomes only distantly related to program actions and components). The use of specific language could help research consumers more readily interpret “mixed effects” in an abstract and regularize the reporting of null or negative results.

Engineering and aviation are also instructive when planning or considering planned obsolescence. Engineers, for instance, collect empirical evidence after mechanical failure events, identifying the system(s) or equipment causing a breakdown in a multipart process. Approximate time to future failure events is estimated, known as the “time-to-failure” or “factor of safety,” or, cheekily, the “factor of ignorance,” because it typically protects the engineer with a wide margin of error (Modarres, 1993; Petroski, 1992; Snook et al., 2003). When the “time to failure” point approaches, relevant systems and equipment are retired and replaced, as a matter of protocol. It may seem grim to assume that an analogous “time to failure” exists in educational programs or policies, but the concepts of effect size durability and decay are relevant here. Cronbach (1975) once warned that a social science experiment is

an observation about recent history, not an enduring conclusion. A decade from now, with changes in the economy, in social morale, in the family structure, and in aspirations, community attitudes will be different. The rules tested in 1970 might compare very differently if tested again in 1985. We tend to speak of a scientific conclusion as if it were eternal, but . . . generalizations decay. (pp. 122–123)

Research is still a time- and context-bound exercise and practitioners should inform education stakeholders that effect sizes may lose potency as a function of time or other forms of change. This could help manage expectations about the success potential of a program or intervention.

Finally, postmortems in some fields provide a template for failure analysis in the social sciences. Post hoc analyses of structural failures, such as the Tacoma Narrows Bridge or the British Comet jet aircraft, are common fare in conventional engineering textbooks. Train and plane accidents are included in the genre of investigative reports. Commercial pilots attend a recurring, closed-door training on human factors and safety that includes a full simulation of recent incidents or accidents and, if possible, video interviews with the personnel involved (CFR 121.427; Code of Federal Regulations, 2020). The latter approach is functional in the sense of fostering candor in disclosure and the psychological safety for learning from colleagues’ mistakes. In medicine, autopsies are considered the “gold standard” to collect data on aging and disease as it relates to failure of the human body (Kuijpers et al., 2014; Shojania & Burton, 2008).

Educational interventions, and the complex systems and subsystems of schools and districts in which they are implemented, are murkier subjects for analysis than a bridge or a cadaver. But there may be identifiable drivers of null results discerned through transparent and orderly analysis. Within education research, some models of this type of analysis exist. The fields of improvement and implementation sciences provide methods to precisely collect data on decisions made at the “practice level” (National Implementation Research Network, 2016) and to identify “local

variabilities” of real-world program implementation that might fail to produce effects (Bryk et al., 2015). Weiss (2002) offers a theory that includes failure mechanisms and describes how a sequencing exercise can assist evaluators in anticipating failure.

METHOD

This chapter intends to build an initial knowledge base on when, how, and why interventions do not reach expectations through a review of published reports of RCTs. Our search focused on recent experimental evaluations of individual programs, instructional materials, classroom technologies, and policies in education, published in reports and journal articles. We aimed to identify evaluations that failed to detect expected effects for one or more main outcomes, as defined earlier. We have not focused on the consequences of a Type I error, that is, false positive findings, or on the probability of Type II error, that is, failure to uncover real intervention effects. These latter will be considered in future research. Studies were included in our sample when the intervention had at least one nonsignificant or negative student-level main effect. Only RCTs were considered. Quasi-experimental studies, regression discontinuity designs, and single-case designs were excluded. The search is limited to education interventions that took place in a K–12 school context, whether during regular school hours or after school.

Our study relies on two important sources of evaluation research, evaluation firms and peer-reviewed journals, that cover a restricted population of RCTs. The resulting assembly of studies is illustrative because there is no population listing from which we can sample. We conducted a search in 2019 of eight firm websites, the WWC website, and three academic peer-reviewed journals, using search engines within organizations’ and publishers’ websites. The goal was to identify studies in this literature with at least one nonsignificant or negative student-level outcome variable, so as to code and analyze the reasons given for an intervention’s full or partial failure to elicit expected outcomes. (See the Online Appendix in the online version of the journal for a full list of organizations, journals, and search terms.) The search yielded 68 total studies, 50 published as evaluation reports and 18 published in peer-reviewed journals. All reports and journal articles were published in the last 10 years (2010–2019). The assembly is illustrative inasmuch as it is not a probability sample from an entire population of such studies.

The pool of studies on which our findings are based consists primarily of large-scale RCTs; the average sample size of a study was 9,000 students or 59 schools. Most studies were cluster-randomized trials that randomly allocated schools to either an intervention or control condition. Approximately four out of five studies (83%) used an academic measure as the main outcome variable of interest. Many studies (64%) used a standardized reading or math test to assess effectiveness. Nonacademic outcome measures were varied, such as change in body mass index or number of completed college applications. The number of main outcome measures ranged from 1 to 13 variables; studies employed five major outcomes, on average. For detail, see the Online Appendix and References.

REPORTED REASONS FOR NULL FINDINGS

The authors of all evaluations in the sample provided reasons for why certain intervention outcomes did not reach expectations. This section presents the reported reasons for null findings. We have discerned four broad types of events that led to null findings in school-based interventions: issues with planning and theory of change (59% of studies), implementation constraints and errors (86%), instability/attrition in the population of participants (41%), and measurement issues (43%). The typology occurs along two dimensions: issues related to the design of the trial (e.g., randomization procedures or data-analytic decisions) and issues related to the design of the intervention (e.g., the theory of change or ease of implementation).

Table 1 presents the four broad categories and the detailed subcategories of reported reasons for null findings. These categories emerged entirely from researchers' stated reasons for or speculations why the intervention did not yield statistically significant results. Resources for this study were not sufficient to corroborate the statements or interview researchers to elaborate on what they had written. Some subcategories overlap conceptually but were kept separate to preserve researchers' specific wording. The subcategories in Table 1 do not sum to overall category percentages because studies often offered multiple reasons for failure in each category.

The following section illustrates how these reported reasons manifest in real experiments and calls attention to common challenges in school interventions that might be addressed or anticipated in the design of a research study.

Planning and Theory of Change

At their best, programs intended to interrupt a status quo in education have an explicit theory of change, a set of propositions about how specific program actions will produce desired outcomes in a population of interest (Newcomer et al., 2015). These propositions are often, but not always, part of the planning or design process initiating a program or intervention. A quantitative theory of change can inform researchers' anticipated program effect size at the outset of a trial, partly by describing the nature of the problem, the size of the population, and the hypothesized way that service delivery will ameliorate or address the problem (Boruch et al., 2019). It is rare to see a theory articulate the role of countervailing forces, such as Weiss's (2002) proposal that theories include "mechanisms that might work for and *against* service delivery" (p. 213).

In total, 41 studies (59%) reported challenges in the planning process that were deleterious to the impact of the intervention. Reported reasons for failure in the planning stages included conflicting goals, no single theory of change, and similar systems already in place, which all relate to the design of the intervention (see Table 1). Approximately 10% of studies stated that their intervention might have failed because it did not include a single vision or theory of change. For example, Greaves et al. (2016) reported that the intervention "Achieve Together" was ineffective because of a "lack of clarity in terms of the initiative's aims and key components affected

TABLE 1
List of Reported Reasons for Null Findings

Events	Reasons Related to Design of Trial	Reasons Related to Design of Intervention
Planning (59%)	Threats to random assignment	Similar systems already in place 20% No single theory of change 10% Multiple roles of teachers/schools 14% Conflicting goals 10% Lack of capacity for change 9%
Implementation (86%)	Intervention period too short Study protocol not fully implemented Follow-up too short Delays in intervention Budget constraints	Incomplete implementation 38% Dose/intensity of intervention too low 28% Difficulties/dissatisfaction with resources 28% Taking time out of teaching content 26% Lack/insufficient training 19% Difficult to deliver 16% Lack of cooperation/buy-in 12% Poor communication/guidance 10% Low student engagement 9% Low teacher engagement 7% Lack of time for cooperation 6%
Instability (41%)	Missing data/low response rate	Attrition of teachers/schools 16% Attrition of students 14% Low student attendance 14% Staffing issues 7%
Measurement (43%)	Issues measuring outcome variables Contamination of comparison Additional variables missing Sample size too small Sample bias	20% 10% 9% 9% 7%

Note. The reported reasons for null or negative findings were identified in each study or report, then coded with the four broad categories. Categories were determined a posteriori, as themes emerged from the data. As distinct subcategories emerged, reasons were additionally coded in subcategories. Finally, subcategories were identified by the authors as relating to the design of the trial versus the design of the intervention.

implementation in so far as staff in case-study schools were unclear about how the program was intended to work” (p. 35). While there was a comprehensive theory of change, it was not clearly communicated to those implementing the intervention.

In most RCTs, interventions are compared to control groups engaged in “business as usual.” In 14 of the studies under review, evaluators suggested that the intervention was too similar to systems already in place in the district or school. In Lindsay et al. (2017), staff survey data showed that intervention and comparison schools “offered students similar supplemental college-readiness supports,” and even then, treatment schools failed to implement those college-readiness components unique to the program (p. 10). In another study, the evaluators reported treatment and control schools were implementing similar systems at the outset, thus reducing the need for a new program (Murphy et al., 2018). In a third study, teachers reported “what they were being asked to do represented little change from what they already do” (Roy et al., 2018, p. 4).

Ginsburg and Smith’s (2016) review of 27 math interventions in the United States notes that research consumers do not often know the extent of service contrast between treatment and control conditions; for example, more than half of studies they reviewed did not provide an adequate description of what the control group was doing. They argue, “Without understanding the comparison’s characteristics, we cannot interpret the intervention’s effectiveness” (p. ii). Similarly, studies in our sample often compared intervention outcomes to a “business as usual” condition that was not clearly defined or documented.

Several evaluators reported they did not fully consider the burden their intervention would place on teachers nor assess the capacity of teachers to implement the intervention. Ten programs were found to demand too much time and commitment from teachers and schools and six programs did not fully consider the necessary preconditions for desired change. The primary strain on implementer capacity was time—such as lack of time to plan new lessons, collect test or outcome data, or introduce new teaching techniques (e.g., Jaciw et al., 2016). In one study, some schools found it difficult to fit intervention sessions, “which varied in length, within the school timetable” (Wigelsworth et al., 2018, p. 4). West et al. (2017) reported that schools in their sample were “overwhelmed by wider performance problems and could not really cope with a further intervention programme, suggesting that the timing of such initiatives needs to be planned in light of the competing priorities a school may be facing” (p. 31). Gersten (2016) described this “painful insight” in a commentary stating that “our priorities were one among many for teachers, and often not the most important” (p. 114). These findings imply that a theory of change is only meaningful if schools have the capacity to implement it.

The most commonly reported issue with the planning process relates to trial design and researcher-practitioner communication: identifying students and schools both eligible to participate in the intervention and capable of carrying out randomization procedures. Nearly 30% of studies in our sample reported difficulties with

either identification or randomization (see Table 1). These challenges significantly reduced evaluators' ability to draw meaningful conclusions and link program activities to desired outcomes. In some cases, the identification and randomization processes were so flawed that evaluators had to change their study design to draw any conclusions.

In Greaves et al. (2016), for example, the authors reported that low levels of recruitment led them to adapt the evaluation design from an RCT to a well-matched comparison group. "The fact that recruitment was challenging is a finding in itself as it may limit the scalability of the programme if schools are unwilling to participate," they concluded (Greaves et al., 2016, p. 48). Other reported challenges included schools deviating from randomization protocols or opting out of the experiment altogether when they were selected for the control group or realized the demands for baseline testing (e.g., Cordray et al., 2013; Husain et al., 2018).

In summary, we suggest that program developers carefully assess the specific schools and districts intended to participate in an intervention and develop a comprehensive, explicit theory of change, with input from practitioner partners, that specifically targets participants' needs. Furthermore, program developers should assess whether the program actions intended to generate outcomes are vulnerable to failure prior to implementation, as suggested by Weiss (2002). This approach will allow us to better understand *ex post* how and why a program succeeded or failed.

Implementation Constraints and Missteps

"It is easier said than done" is the sentiment of many evaluators describing challenges with program implementation. Eighty-six percent of studies reported at least one major problem during implementation, and some as many as nine problems. The most common reasons given for failure as a function of implementation included the intervention period not being long enough, the dose or quality of intervention not being high enough, lack of training, and difficulties with program resources and/or logistics. The 16 subcategories of implementation issues listed in Table 1 speak to the large variety of issues faced by students, teachers, school leaders, and program staff during implementation.

Student engagement is a driving factor of any successful intervention. If students are not motivated, interested, or engaged in a school-based activity, it will be difficult to change their performance, behavior, or mindset as a consequence. Five studies reported difficulties with student engagement, such as students not participating in lessons or not completing assignments. For instance, Gonzalez et al. (2018) found that more than one third of students assigned to the treatment group did not engage with the reading intervention. Sometimes lack of participation was an intentional student or family decision; in one case teachers terminated the intervention under study because students resented leaving a creative writing class to participate in it.

Lack of teacher engagement or involvement can also constrain implementation. Wigelsworth et al. (2018) examined the "FRIENDS for life" intervention

implemented in 79 U.K. primary schools, aimed at promoting emotional resilience and preventing student anxiety and depression. It was delivered by a team of external project officers assigned to caseloads of schools. The authors aver that one drawback of this model “could be the lack of reinforcement of the programme throughout the school life” (p. 46), as teachers were not closely involved with the intervention. In addition, program officers had a difficult time to find quiet spaces for their school-based activities, which could signal poor communication or low school commitment. Last, project officers were less familiar with students’ backgrounds, troubling emotional connections. Greater teacher involvement in implementation processes may have alleviated some of these problems. Studies did not often distinguish whether engagement issues (student or teacher) were attributable to participants’ predispositions before the intervention or to the design of the intervention itself.

It is no surprise that time and timing is a major factor in deploying any large-scale intervention. There are challenges for those not familiar “with the bell-schedule nature” of the school context (Gersten, 2016, p. 114). For example, teachers can only participate in full-day training sessions during the summer or institute days, and training delays can jeopardize the implementation of a full intervention. Many studies reported that the intervention period was either too long or too short. Some programs that targeted student motivation and engagement were found to have negative impacts if they were too long (e.g., Jayanthi et al., 2017), while programs that targeted students’ academic attainment were often found to be too short (e.g., Jerrim et al., 2015). Foorman (2016) argues that many researchers must “learn to adjust their view of ideal implementation to the reality of school life” (p. 9), especially when it comes to time.

The most common timing challenge occurred when a classroom-based intervention was viewed as additional to ordinary teaching duties. For instance, Eddy et al. (2010) found that very few teachers who were asked to implement a five-unit reading program completed the program. They offer three reasons:

1) obligations to cover other material in class such as novels or standardized test preparation; 2) dissatisfaction with some elements of the program [. . .]; and 3) desire to integrate other sources of literature or language arts instruction that was not textbook based. (Eddy et al., 2010, p. 19)

Teachers also reported that the recommended pacing guide was unrealistic. This example illustrates how important it is to clarify whether an intervention is additive to, or in lieu of, other teaching duties and develop realistic pacing guides for successful implementation.

Many of the reasons for a null or negative outcome discussed in this section are linked to the design of the intervention. The most common reasons reported in Table 1 are incomplete implementation or an intervention implemented with too low of dosage or intensity. Studies that included fidelity of implementation ratings typically found differences in implementation across classrooms and associated a low dose of the intervention (e.g., teachers did not participate in all training sessions or did not

complete all required lessons) with lower intervention effects. Often teachers reported low-dose implementation due to difficulties with timing (see above) or with program delivery, such as not having required resources. For instance, two different programs that aimed to integrate technology in the classroom (Jaciw et al., 2012; Worth et al., 2017) reported challenges with timely equipment delivery, software updates, and stable Internet connections. Both reports echoed that the “effect was dependent on local conditions” (Jaciw et al., 2012, p. ii). These discrepancies between the design and implementation are sometimes described as “program drift” and can drastically decrease the benefits of a program (Vaden-Kiernan et al., 2018).

To better understand implementation quality, researchers are calling for a more careful measurement of implementation indicators, “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” (Jacob et al., 2019, p. 586). They point out that these factors should be considered in the early design of an intervention. We also believe more can be learned from an intervention with a null or negative outcome if evaluators gather detailed implementation data throughout the experiment and link particular aspects of implementation, such as student engagement, intervention timing, teacher buy-in, or study design, to the success or failure of the trial.

Instability in the System

Schools are changeable environments. Students move in and out. Teachers are reassigned to different subjects or grades, or leave the school or profession. New academic standards are adopted. Large-scale assessment vendors change, as do measured knowledge and skills. Schools close or “turnaround,” site, district, and state leadership roles turn over, and priorities often shift as a result. Instability during a trial, be it teacher attrition or student mobility, can undercut the training or knowledge imparted by the tested intervention and shrink the number of people who receive the intended dose. And the departure of teachers or entire schools from an analytic sample can reduce study power, particularly in cluster randomized trials. Forty-one percent of studies in this sample attributed a failed outcome to instability, including factors like student or teacher attrition, low participant attendance, other forms of nonresponse, and staffing issues (Table 1).

The loss of students, teachers, leaders, or entire schools from a study accounted for failure in approximately one quarter of the evaluation reports in our sample. Many RCTs presumed the relative stability of their subjects. Yet empirical benchmarks based on longitudinal surveys place average national student mobility rates from 12% for a kindergarten to Grade 1 transition to 46% for a Grade 1 to Grade 5 transition, “[implying] that a study designed to follow students from Grade 1 to Grade 5 should anticipate losing half of the baseline sample if the study collected data only on students who remained in the study schools” (Rickles et al., 2018, p. 632).

High levels of instability can dilute the effectiveness of an intervention and reduce the statistical power of a study and the probability of detecting a true effect (Boruch

et al., 2016). Five studies in our assembly reported a loss of one quarter to one half of participating students, sometimes as a result of losing several schools in a cluster sample. For example, an evaluation of a foreign language learning approach in primary schools (Wiggins, Parrao, et al., 2017) lost 12 schools (28% of all classrooms) following random allocation before the start of the intervention and another nine schools between the intervention's start and posttest data collection.

The reasons offered for school and classroom dropouts ranged from logistical difficulties (e.g., schools could not release teachers for mandatory training or could not configure classes as required by the intervention) to midyear personnel turnover (e.g., staff who were trained on the intervention then left the role for professional reasons or went on long-term leave). In other cases, schools suddenly stopped responding to researchers. Pane et al. (2010), evaluating the efficacy of a technology-based geometry curriculum, expressed interest in the treatment effect as a function of teachers' prior experience with the curriculum, but found that only 3 of 19 teachers in the sample used the curriculum for all 3 years of the study; most used it for 1 year. These examples illustrate known instabilities that, to a certain extent, can be anticipated. Knowing that school environments are not stable, interventions and trialists might seek to detect smaller effects from a reduced share of the target population, and the trial might better forecast an intervention's effects as implemented.

Most trials experiencing detrimental student attrition raised the issue of nonresponses for outcome data. A few studies reported nonresponse issues because there were gaps, omissions, and errors in extant administrative records rather than students actually leaving the study. For example, Lindsay et al.'s (2017) evaluation of a college readiness support program reported 36% of all student subjects were missing a baseline Grade 8 standardized math test score in records provided by the participating schools and districts. Missing data are common, especially when collected during a different stage of schooling than the intervention or collected principally for a purpose unconnected to the study (Rickles et al., 2018). To offset these limitations, Rickles et al. (2018) suggest that researchers explore whether the time and effort of a wider data collection strategy for extant information is warranted (such as collecting statewide standardized achievement test records in lieu of school- or district-wide) to reduce nonresponse rates.

The related issue of null or negative outcomes as a function of low participant attendance or engagement was raised in 23% of studies in the sample. Sometimes one stakeholder group's low rates of involvement affected others. For example, Cavalluzzo et al. (2012) evaluated a hybrid instructional model that combined online and face-to-face instruction in Algebra I and rated 30% of teachers as having "low engagement" in professional development activities for the project. This likely contributed to the 65% of treatment students measured as having relatively low levels of engagement in online learning. Similarly, Gonzalez et al. (2018), evaluating a literacy intervention for struggling adolescent readers, found that two thirds of students assigned to the treatment received fewer than the recommended number of sessions, partly because only 2 of 10 study schools offered sessions daily.

Variation in levels of teacher and student engagement can be stark. In a literacy professional development program, Jaciw et al. (2016) found that teachers who met the criteria for implementation fidelity attended eight of nine possible sessions during the school year, on average, versus the rest of participating teachers, who attended only one of nine sessions, on average. In a multiyear summer school intervention, Augustine et al. (2016) reported that nearly half of students did not attend a single day in the second summer, compared with a smaller subset of students who reported consistently high attendance in the first and second summers. In other cases, such as Gonzalez et al.'s (2018) evaluation of a literacy intervention, attendance differed significantly by grade.

Finally, staffing issues were reported in some studies (less than 10%) as a contributing factor in null or negative trial results. School districts typically receive notification of teachers' intent to leave or transfer at the end of a school year and fill the bulk of teacher vacancies during the summer months (Levin & Quinn, 2003; Papay & Kraft, 2016). Thus, interventions that commence with mandatory summer activities or training for teachers, but do not attend to whether teacher assignments are finalized for the next academic year, invite discontinuity in implementation. For example, in Bos et al. (2012), it was a simple fact that "districts did not know in June which teachers would be employed in which grade level the following September" (p. 48). Multiyear trials, such as Wiggins, Sawtell, and Jerrim's (2017) evaluation of teacher and student handheld electronic devices that spanned 2 years, leave programs vulnerable to annual staffing changes.

One form of instability seldom mentioned in the sample of reports was the threat of trial sites abandoning the intervention for a new product or policy. In theory, an intervention's perceived tenure could affect participant buy-in. Public school district procurement for products like curriculum, education software, classroom technologies, and teacher professional development is a Byzantine process in both the United States and the United Kingdom (Maas & Lake, 2015; Younie, 2006). Although the average vendor contract length in a U.S. public school district is unknown, products such as curriculum are adopted on cycles as short as every 3 years (Zinth, 2005). Augustine et al. (2018), evaluating a Restorative Practices intervention to reduce suspensions and improve school climate and culture, and finding mixed effects, included a rare teacher perspective on the issue of perceived intervention tenure:

If they [the district] don't provide schools with support, Restorative Practices will disappear. This is what happens in the district. They give us something and they either don't give us everything we need to fully implement it, or they back off [in their level of support for an initiative] and staff stop doing it because something new (another initiative) takes its place. . . . If the district doesn't support these schools, Restorative Practices will not be around in these schools in five years. (p. 57)

Measurement Issues

To assess the impact of an intervention, an evaluator has to decide what, when, and how to measure change, and how many changes one might expect as a result of

treatment. On average, studies in the sample collected data for five outcome measures, with a range of 1 to 13 measured outcomes. Many studies seemed to have focused their data collection on the major, final program outcomes, such as students' summative academic performance, without plans to measure intermediary outcomes like participation, motivation, and engagement during implementation. It is difficult to collect data on intermediary outcomes retrospectively because it takes time and resources for both evaluators and program participants and often relies on self-reported recollections.

Several evaluators identified intermediary and final outcome variables which should have been added to the design of the trial when discussing null or negative effects. Five studies specifically noted outcome variables that would have shed more light on intervention impacts, and 20% of studies stated they had problems with their outcome variable as it did not fully capture the objective of the intervention. Common problems included the use of standardized reading assessments to measure the impact of a diverse array of learning interventions; dependence on single performance measure to assess program outcomes; and overreliance on self-administered surveys.

For example, the evaluation of an extracurricular reading initiative, "Chatterbooks," which aimed to increase reading motivation by providing students with books to read for pleasure (Styles et al., 2014) relied on a standardized reading test administered immediately after the intervention and again 3 months later. The test of reading achievement did not adequately address the research objective, a child's motivation to read. In another reading intervention, economically disadvantaged students were shipped books to read over the summer, matched to their reading level and interests. The program found no statistically significant impacts on reading comprehension, as measured by a standardized assessment tool. However, the evaluators noted that no implementation or follow-up data were collected to measure whether participating students received the matched books or read some or all of them (Wilkins et al., 2012). The study could not differentiate between students in the treatment group who did or did not read the intervention's books, a major limitation in the interpretation of the null findings.

Several studies were prevented from finding significant effect sizes due to statistical power (e.g., Husain et al., 2018). For instance, Arens et al. (2012) point out that their intervention was not powered to detect effect sizes smaller than 0.35 standard deviations and they used only one outcome measure. Considering effect sizes as small as 0.20 or 0.30 are generally considered practically meaningful (Spybrook & Raudenbush, 2009), and some researchers allege that "small and probably heterogeneous effects" are "the facts" of education research (Valentine, 2019, p. 612), the power of Arens et al.'s study is a limitation. It could be avoided by using more than one outcome measure, increasing sample size, or preventing attrition, as reductions in sample size also reduce power (Spybrook & Raudenbush, 2009).

The last set of measurement issues to be discussed here relate to the study sample. The reasons reported for failure as it relates to the sample are commonly discussed in the evaluation literature, so we mention them only briefly. Seven studies reported that the contamination of students and the consequent dilution of program effects negatively affected study outcomes. Five of seven studies that reported problems with contamination had randomly assigned participants (usually teachers) at the classroom level within the same school. Randomization of program participants at the individual or classroom level appeared to be more problematic than randomization at the school level. If randomization at the school level is not possible, researchers can attempt to reduce contamination by asking teachers not to share program resources with coworkers until the end of the intervention (e.g., Cordray et al., 2013) or offering other incentives.

Five studies in our review reported that attrition and sample selection problems might have led to sample bias. It is common practice in most evaluation reports (e.g., Heller, 2012) to conduct statistical tests that rule out forms of bias associated with attrition, such as tests revealing whether differential attrition led to nonequivalence of baseline characteristics of treatment and control group members. The evaluators at the Education Endowment Foundation (EEF) in the United Kingdom recommend strategies to minimize the loss of sample participants or data, such as reducing the frequency of student testing, which can put an extra burden on schools, or carefully considering the timing of test data collection and using extant, national student databases to measure key outcomes (Martin et al., 2018). The EEF argues that measurement attrition has been identified as the single biggest issue affecting the security of trials and reducing confidence in findings (A. Dawson et al., 2018). The WWC in the United States also provides a detailed “Attrition Standard” for researchers navigating the relationship between attrition, power, and bias (WWC, 2015).

SUMMARY AND IMPLICATIONS

We have set out the reasons for failure to detect expected effects that were offered by researchers who led large-scale education RCTs conducted over the last 10 years. Our work identifies a typology of four failure mechanisms based on the empirical study of nearly 70 RCTs and therefore extends existing literature on the nature and consequence of null findings. The literature cumulatively suggests a number of ways researchers can guard against null results, including articulating an intervention’s logic model; employing a small number of directly program-relevant outcome measures; acknowledging that realistic effect sizes in school settings are often small; and collecting more data on fidelity of implementation than convention stipulates (e.g., C. J. Hill, 2019; H. C. Hill & Erickson, 2019; Jacob et al., 2019; Kim, 2019; Valentine, 2019). Based on our findings, we provide additional and augmented implications for researchers below.

Implications: Design Stage

Existing recommendations to advance the quality of evidence on interventions and prevent experimental failure primarily focus on the design of a trial rather than the design of the intervention being tested. For example, Song and Herman (2010) provide a reference guide for evaluators to improve education RCTs based on the first phase of the WWC's study reviews, in which only 5% of the 1,500 reviewed studies met rigorous WWC standards. They advise researchers to attend to sampling design, power analysis, construct validity and related measurement issues, implementation fidelity, and critical data analysis decisions. Little mention is made of the interventions themselves, though. For instance, in a section on implementation, the authors discuss issues related to the validity and reliability of outcome variables, the importance of collecting implementation data, and issues arising from attrition, but neglect the importance of a comprehensive theory of change or the role of teacher engagement.

Similar to the work done at the WWC in the United States, the Early Intervention Foundation (EIF) reviews the quality of education RCTs in the United Kingdom and provides recommendations for evaluators to improve the quality of study designs. A recent EIF report (Martin et al., 2018) discusses six common evaluation pitfalls surfaced by the Foundation's work, issues that "frequently reduce our confidence in study results, culminating in a lower-than-expected evidence rating" (p. 7). Their discussion focuses on attrition, study participant exclusion, fair comparison groups, measurement issues, and small sample sizes, which all relate to trial design.

Our analysis demonstrates that both the design of the intervention and design of the trial are critical dimensions to consider when trying to improve the quality of evidence from RCTs. Recommendations and checklists produced by organizations such as the WWC and EIF focus on trial design and address the researcher or evaluator as a solitary agent, rather than a collaborator with practitioners and educators, who have unique insight into both trial and intervention issues. We suggest expanding checklists for evaluators to include important aspects of intervention design and propose practitioner input in most if not all areas of trial *and* intervention design.

For example, educators in intervention schools should be involved in articulating a theory of change that they understand and support for their student population. Ming and Goldenberg (2021, this volume) note that quality research must align its topic to practical priorities; alignment is unlikely without collaboration during design stages. Moreover, evaluators should plan regular implementation check-in meetings with educators and site leaders to detect issues with resources early on (see further recommendations in the checklist in Online Appendix A). Sixty percent of reported reasons for failure in this chapter relate to the design of the intervention and more can and should be done to anticipate and overcome challenges in this regard.

Implications: Implementation Stage

As noted, most studies in the sample (86%) reported at least one major problem during implementation, such as a lack of adequate time allotted for the intervention

or varying levels of participant dosage. However, the character and collection of implementation data varied greatly. This is consistent with other research on the relationship between implementation fidelity and null results, such as H. C. Hill and Erickson (2019), who concluded, “whereas conventional wisdom in policy analysis often locates null results in implementation failure, we have no estimates of the extent to which this is true, particularly in recent, rigorous trials of educational interventions” (p. 590). Detailed implementation data would verify researchers’ claims about why and how an intervention did not work, moving evidence from speculation to empiricism.

Models for collecting implementation data exist, such as the resources at the National Implementation Research Network at the University of North Carolina. Our analysis suggests that much more could be done to diagnose when and why a program did not work as a function of its adherence to implementation fidelity guidelines, using data collected throughout the experiment. Such evidence collection could also capture “productive adaptations” made by practitioners to improve upon the program model (McLaughlin & Mitra, 2001) and lead to a deeper understanding about program elements that are essential as designed, versus those elements that can be adapted or localized without compromising outcomes (Kim, 2019).

One way to ensure the collection of adequate implementation data is the advancement of theory-based evaluations (TBEs). TBEs advocate for the collection of implementation data to “follow the chains of program theory” (Weiss, 1997) and draw more useful, relevant, and practical conclusions when a program fails. Evaluators collect data for each step of an intervention (preliminary, intermediary, and long-term) to evaluate the phased sequences of actions which are meant to drive program outcomes. If the sequence breaks down, the evaluator can tell at what point the breakdown occurred and modify the intervention. As our analysis demonstrates, such detail is not often explicit in a theory of change for either evaluators and implementers (i.e., educators). This may dilute the effectiveness of an intervention, or at least hinder a nuanced understanding of effects.

In several studies, educators pointed out in postimplementation interviews that they were unsure how the program they were asked to implement was intended to work (e.g., Greaves et al., 2016). Another related way to improve implementation data collection is to increase practitioner investment in this stage, by sharing the data on program adherence with educators and stakeholders in an ongoing way or soliciting and using their formative feedback. The improvements to research quality that arise from successful research-practitioner partnerships are well-documented in this volume (e.g., Crain-Dorough & Elder, 2021, this volume; Welsh, 2021, this volume).

Implications: Anticipating Instability

Studying failure could inform future program, product, and policy designs to avoid or accommodate known instabilities in educational settings. Rather than classify the departure of one third or one half of teachers or students engaged in a

multi-year trial as a limitation post hoc, researchers might anticipate attrition as a known factor and define success more realistically. Power analysis, for example, could incorporate empirical estimates of student and teacher mobility. Recent work by Rickles et al. (2018) amends the standard formulas for minimum detectable effect sizes, adding parameters for overall and differential student attrition rates. The authors also calculate empirical benchmarks for student attrition during different grade transition periods using nationally representative, longitudinal surveys.

For teacher attrition, Taylor and West (2020) provide probability estimates that researchers can similarly use in power analyses. Estimates of one- and five-year teacher mobility rates at state and major city levels are also available from analyses of longitudinal administrative records on entire teacher workforces (e.g., Allensworth et al., 2009; Boyd et al., 2002; Chao et al., 2016; Frisone et al., 2016; Lankford et al., 2002; Papay et al., 2017; Rayes et al., 2016; Ye et al., 2016). Much of the retention and attrition literature highlights how first-year or early career teachers (who are the focus of many interventions, such as teacher professional development programs) are more likely to leave their jobs. Mobility estimates that depend on public use records available from states could be used for more precise power analysis and sample size estimation before an experiment, or for design considerations, such as limiting the amount of training required in settings with high teacher churn.

Our analysis also suggests that researchers might approach trials with more realistic, empirically based expectations about what can be accomplished in the complex system of a school. Understanding expected staff and student turnover and other forms of planned change at the site level can help a researcher assess whether the proposed minimum detectable effect size and clinically meaningful outcomes are sensible. For example, in large-scale educational RCTs, Jacob et al. (2019, p. 586) point out that “rarely is there a no services option” because business as usual is some variant of the intervention itself, such as a math intervention compared to regular math instruction. Given that control students’ services so often overlap with treatment students’ services, and that student and teacher mobility and absence are known and somewhat predictable aspects of schools, a credible effect size may be less than the 0.20 MDES studies are typically powered to detect (Jacob et al., 2019; Spybrook & Raudenbush, 2009; Valentine, 2019).

Implications: Success Criteria

Testing many outcome variables can make research vulnerable to detecting a significant result by chance, even when no effect is present. Moreover, testing multiple outcomes without specifying or prioritizing those most likely to change as a direct consequence of the intervention (vs. secondary or distal outcomes) can render it difficult to determine whether an intervention failed or is worthy of replication. In this assembly of experiments, the studies in which all major outcomes were statistically nonsignificant reported fewer outcome measures (3.5, on average) than studies in which all major outcomes were mixed (positive, negative, and/or null) or negative

(10.2 outcome measures, on average). One implication is that a clearer picture of intervention performance emerges from fewer outcome measures.

C. J. Hill's (2019) recent commentary on null results posits that the education field would generate better knowledge if researchers

[limited] the number of confirmatory outcomes and the number of confirmatory subgroups examined, [and,] if additional exploratory outcomes and subgroups are examined, it would . . . require committing to consistently reporting and summarizing those results as exploratory—with emphasis on understanding imprecisely estimated null results and mixed results, not just those that are statistically significantly different than zero. (p. 608)

Supplementing site- or subgroup-level treatment effects with descriptive analyses would help researchers explore whether reasons for null or negative main effects apply to all study subjects. Limiting the number of subgroups may assist studies in better preparing to detect such effects. Few RCTs funded by IES are powered to detect heterogeneous treatment effects across groups (Spybrook, 2014) and, when powered, are much more likely to reveal clinically meaningful student-level moderator effects than school- or teacher-level moderator effects (Spybrook et al., 2020).

Education researchers would benefit from more clearly articulated best practices to understand and report on impact heterogeneity or imprecisely estimated results. Convention in other fields could be instructive here, such as tests for equivalence common in the pharmaceutical industry, which demonstrate that a small, nonzero difference between a treatment and control condition is equivalent or small enough to ignore (Lakens et al., 2018). These ideas are not new. Cronbach (1975) argued for the value of descriptive and exploratory information in the event of statistically ambiguous results: “descriptions encourage us to think constructively about results from quasi-replications, whereas the dichotomy significant/nonsignificant implies only a hopeless inconsistency” (p. 124).

Implications: Postmortems

As suggested earlier, the practice of conducting a post-facto analysis of incidences of failure is common in some scientific domains. Indeed, the practice of reporting failure conscientiously has become so common that the *Journal of Negative Results in Biomedicine* has ceased to publish because, according to its web site, the journal “is no longer needed.”

To judge from the assembly of studies reviewed here, however, orderly and transparent postmortems are not yet common in education. The few models and methods at hand can nonetheless be instructive. It seems reasonable to recognize and try to improve them. Consider two examples which do not come from our sample. The first concerns a community-based violence prevention program, and the second concerns a massive professional development program for teachers in Italy.

Williams and Mattson's (2006) paper is largely based on exit focus group discussions conducted with participating youth, program coordinators, and researchers following a

trial that led to insubstantial effects. The researchers tape-recorded discussions; transcripts served as the main source of data for reporting on null results. The authors were able to identify youth characteristics that are linked to more positive outcomes and determine flaws in the study design. Most important, they reported that despite the challenges the program coordinators faced in implementing the program and the disappointing outcomes of the study, the evaluation team was successful in providing a platform for learning for all stakeholders involved which will promote future collaboration.

The Italian M@t.uel Program was tested in a cluster randomized trial involving 175 schools, over 660 teachers, and 11,000 students. This professional development intervention had no substantial effect on the main outcome—lower secondary school students' scores on standardized tests of math achievement. The researchers' post mortem on possible and plausible reasons for failure depended on posttrial structured CATI interviews with teachers, teacher journals diaries kept during the trial, and administrative records on the pipeline of teachers and its leakage during all stages of the trial (Abbiati et al., 2014). So, for instance, they explicate eight or so categories of reasons that teachers gave for leaving the course or not turning up and a half dozen categories of levels of noncompliance. Their tables and qualitative material are instructive. They explain incentives and their absence, voluntary versus principal directed decisions, and sheer burden of travel and related factors as influential.

Each of these two examples, and the variability in reporting in the studies reviewed here, invite one to think about standardizing checklists for plausible reasons for failure, using such checklists in postmortems, and including completed checklists in publications. Checklists are fundamental in assuring transparency and uniformity in reporting. They may also help one learn from others about anticipating missteps and mistakes in developing and executing trials.

Implications: Incentives

Incentives for cooperating in a trial, the lack thereof, or their weakness are rarely considered deeply in authors' reports on the plausible reasons for a failure to detect expected effects in trials. About 14 of the reports in our sample do say that incentives were part of the experiment's design, for example, control group teachers being offered a gift card, schools offered money, students offered ice cream socials, and so on. But only two of these reports return to the topic in their discussion of less than expected effects. The exceptions include Augustine et al.'s (2016) remarking that the incentives offered in a test of a summer program appeared to not work in the face of parents' relaxed attitudes about such a program. In the other exception, Murphy et al. (2018) declare that "the absence of external observers [of teachers] and incentives in our program may explain the contrast of these results with . . . work which shows a positive influence of teacher observation and feedback on pupil outcomes" (p. i).

This relative inattention to incentives is puzzling on several accounts. Research on survey methods that aim to achieve high response rates, for example, is studded with tests of different incentive types and levels for enhancing cooperation. Because

randomized trials can be considered as a project involving two or more parallel surveys of the groups being compared, the survey RCT research on incentives by Lavrakas et al. (2019) is pertinent. Economists' research on incentives for teachers in the context of enhancing children's achievement seems relevant but to judge from Hanushek (1996) they are "seldom evaluated in any systematic manner" (p. 43), and they appear to be rarely, if ever, considered explicitly in the context of education trials.

To be sure, disincentives (as opposed to positive incentives) that are arguably related to failure of implementation and other problems are identified at times in the reports that we reviewed. The time and resource demands placed on teachers in a trial can be construed as disincentives and are recognized as such. Nonetheless, positive incentives that offset or attenuate the effect of the disincentives are not considered explicitly in any theory of change that we have seen in reports. See also the section on postmortems above.

The implication is that incentives and disincentives can be and should be part of an intervention's design. Moreover, the statistical designers of a controlled trial, including technical aspects of randomization and measurement, might do well to take incentives for cooperation into account as colleagues in the survey methods arena do.

CONCLUSION

Teachers in schools know that learners can greatly benefit from making mistakes ("productive failure," in classroom parlance). The education research community is encouraged to more actively learn from mistakes or missteps when implementing large-scale RCTs that do not elicit expected effects, for example, conducting postmortems or collecting rigorous implementation data. The framework introduced here for analyzing null and negative outcomes aims to augment researchers' understanding of the conditions in which educational interventions succeed or fail. We propose that failures in education can be studied in a systematic and transparent way to maximize the evidence yielded from studies.

RCTs with at least one nonsignificant or negative major outcome are an untapped source of high-quality evidence on major sources of failure that can be preempted through thoughtful design, statistics, or both. The systematic study of failure in education sciences, as in other disciplines, can enrich our understanding. Most important, it is to improve the process of education and the well-being of children.

ACKNOWLEDGMENTS

We are grateful for support of early work on the topic through a grant from the National Science Foundation, DR 1337237.

ORCID iD

Claire Allen-Platt  <https://orcid.org/0000-0002-6344-4579>

REFERENCES

- Abbiati, G., Argentin, G., & Pennisi, A. (2014). Learning from implementation: The case of the evaluation of a professional development programme for mathematics teachers in Italy. In S. Kalliola (Ed.), *Evaluation as a tool for research, learning, and making things better* (pp. 223–240). Cambridge Scholars Publishing.
- Achieve. (2018). *Proficient vs. prepared 2018: Disparities between state tests and the 2017 National Assessment of Educational Progress (NAEP)*. https://www.achieve.org/files/Proficient%20vs.%20Prepared%20May2018_1.pdf
- Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago public schools*. Consortium on Chicago School Research at the University of Chicago.
- Alvarez & Marsal Holdings LLC. (2018, January 26). *Final report: District of Columbia Public Schools Audit and Investigation* (Contract Number # CW57247). https://osse.dc.gov/sites/default/files/dc/sites/osse/release_content/attachments/Report%20on%20DCPS%20Graduation%20and%20Attendance%20Outcomes%20-%20Alvarez%26Marsal.pdf
- Arens, S., Stoker, G., Barker, J., Shebby, S., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2012). *Effects of curriculum and teacher professional development on the language proficiency of elementary English language learner students in the central region* (NCEE 2012-4013). Mid-continent Research for Education and Learning.
- Augustine, C., Engberg, J., Grimm, G., Lee, E., Wang, E., Christianson, K., & Joseph, A. A. (2018). *Can restorative practices improve school climate and curb suspensions? An evaluation of the impact of restorative practices in a mid-sized urban school district*. RAND Corporation.
- Augustine, C. H., McCombs, J. S., Pane, J. F., Schwartz, H. L., Schweig, J., McEachin, A., & Siler-Evans, K. (2016). *Learning from summer: Effects of voluntary summer learning programs on low-income urban youth*. RAND Corporation.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543–554. <https://doi.org/10.1177/1745691612459060>
- Bandeira de Mello, V., Rahman, T., Fox, M. A., & Ji, C. S. (2019). *Mapping state proficiency standards onto NAEP scales: Results from the 2017 NAEP reading and mathematics assessments* (NCES 2019-040). Institute of Education Sciences, National Center for Education Statistics.
- Boruch, R., Allen-Platt, C., & Gerstner, C. (2019). To randomize or not to randomize? That is the question. *New Directions in Evaluation*, 2019(163), 73–82. <https://doi.org/10.1002/ev.20373>
- Boruch, R., Merlino, J., Bowden, J., Baker, J., & Chao, J. (2016). *In search of terra firma: Administrative records on teachers' positional instability across subjects, grades, and schools and the implications for deploying randomized controlled trials*. https://repository.upenn.edu/gse_pubs/393/
- Boruch, R., & Ruby, A. (2015). To flop is human: Inventing better scientific approaches to anticipating failure. In *Emerging Trends in the Social and Behavioral Sciences*. <https://doi.org/10.1002/9781118900772.etrds0362>
- Bos, J., Sanchez, R., Tseng, F., Rayyes, N., Ortiz, L., & Sinicrope, C. (2012). *Evaluation of quality teaching for English learners (QTEL) professional development* (NCEE 2012-4005). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2002). *Initial matches, transfers, and quits: Career decisions and the disparities in average-age teacher qualifications across schools* (Stanford CEPA Working Paper). https://cepa.stanford.edu/sites/default/files/Initial_Matches_Transfers_and_Quits.pdf

- Bryk, A. S., Gomez, L., Grunow, A., & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Publishing.
- Bureau of Transportation Statistics. (2020). *Glossary*. <https://www.transtats.bts.gov/glossary.asp>
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.
- Cavalluzzo, L., Lowther, D. L., Mokher, C., & Fan, X. (2012). *Effects of the Kentucky Virtual Schools' hybrid program for algebra I on grade 9 student math achievement* (NCEE 2012-4020). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Chao, J., Park, J., & Boruch, R. (2016). *Ambient positional instability among Illinois teachers, AY 2007–2012: A briefing* (CRESP Working Paper/Briefing). https://repository.upenn.edu/gse_pubs/395
- 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>
- Code of Federal Regulations. (2020). *Recurrent training*. https://www.ecfr.gov/cgi-bin/text-idx?SID=47e3db32559d6354948f7080e1960aa4&mc=true&node=pt14.3.121&rgn=div5#se14.3.121_1427
- Cordray, D. S., Pion, G. M., Brandt, C., & Molefe, A. (2013). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement* (NCEE 2013-4000). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Core Standards. (2019). *About the standards: Development process*. <http://www.corestandards.org/about-the-standards/development-process/>
- Crain-Dorough, M., & Elder, A. C. (2021). Absorptive capacity as a means of understanding and addressing the disconnects between research and practice. *Review of Research in Education*, 45(1), 67–100. <https://doi.org/10.3102/0091732X21990614>
- Cronbach, L. (1975). Between the two disciplines of scientific psychology. *American Psychologist*, 30(2), 116–127. <https://doi.org/10.1037/h0076829>
- Dahlin, K. B., Chuang, Y., & Roulet, T. J. (2018). Opportunity, motivation, and ability to learn from failures and errors: Review, synthesis, and ways to move forward. *Academy of Management Annals*, 12(1), 252–277. <https://doi.org/10.5465/annals.2016.0049>
- Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: Reflections from England's Education Endowment Foundation. *Educational Research*, 60(3), 292–310. <https://doi.org/10.1080/00131881.2018.1500079>
- Dawson, P., & Dawson, S. L. (2018). Sharing successes and hiding failures: “Reporting bias” in learning and teaching research. *Studies in Higher Education*, 43(8), 1405–1416. <https://doi.org/10.1080/03075079.2016.1258052>
- Eddy, R. M., Ruitman, H. T., Hankel, N., & Sloper, M. (2010). *The effects of Pearson Prentice Hall literature (2010) on student performance: Efficacy study*. Cobblestone Applied Research and Evaluation, Inc.
- Foorman, B. (2016). Introduction to the special issue: Challenges and solutions to implementing effective reading intervention in schools. *New Directions for Child and Adolescent Development*, 2016(154), 7–10. <https://doi.org/10.1002/cad.20172>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Public bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505. <https://doi.org/10.1126/science.1255484>

- Frisone, M., Hooks, T., Ye, T., & Boruch, R. (2016). *Ambient positional instability among core subject Arkansas public school teachers: Interim report*. https://repository.upenn.edu/gse_pubs/394
- Gersten, R. (2016). Commentary: The tyranny of time and the reality principle. *New Directions for Child and Adolescent Development*, 2016(154), 113–116. <https://doi.org/10.1002/cad.20171>
- Ginsburg, A., & Smith, M. (2016). *Do randomized controlled trials meet the “gold standard”?* American Enterprise Institute.
- Gonzalez, N., Macintyre, S., & Beccar-Varela, P. (2018). *Challenges in adolescent reading intervention: Evidence from a randomized control trial*. Mathematica Policy Research.
- Greaves, E., Sianesi, B., Sibieta, L., Amin-Smith, N., Callanan, M., & Hudson, R. (2016). *Achieve together: Evaluation report and executive summary*. Education Endowment Foundation.
- Hall, L. S., & Callahan, D. (2018). *It's one of the biggest failures yet in K–12 philanthropy. What are the lessons?* <https://www.insidephilanthropy.com/home/2018/7/54/another-lesson-in-k-12-philanthropy-the-gates-teacher-effectiveness-initiative>
- Hanushek, E. A. (1996). Outcomes, costs and incentives in schools. In *Improving America's schools: The role of incentives* (pp. 29–52). National Research Council.
- Heller, J. I. (2012). *Effects of Making Sense of SCIENCE™ professional development on the achievement of middle school students, including English language learners* (NCEE 2012-4002). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Herrington, C. D., & Maynard, R. (Eds.). (2019). Randomized controlled trials meet the real world: The nature and consequence of null findings. *Educational Researcher*, 48(9), 577–579. <https://doi.org/10.3102/0013189X19891441>
- Hill, C. J. (2019) Commentary on the null results special issue. *Educational Researcher*, 48(9), 608–610. <https://doi.org/10.3102/0013189X19891432>
- Hill, H. C., & Erickson, A. (2019). Using implementation fidelity to aid in interpreting program impacts: A brief review. *Educational Researcher*, 48(9), 590–598. <https://doi.org/10.3102/0013189X19891436>
- Husain, F., Wishart, R., Marshall, L., Frankenberg, S., Bussard, L., Chidley, S., Hudson, R., Votjkova, M., & Morris, S. (2018). *Family skills: Evaluation report and executive summary*. Education Endowment Foundation.
- Jaciw, A., Toby, M., Ma, B., Lai, G., & Lin, L. (2012). *Measuring the average impact of an iPad algebra program*. Empirical Education Inc.
- Jaciw, A. P., Schellinger, A. M., Lin, L., Zacamy, J., & Toby, M. (2016). *Effectiveness of Internet-Based Reading Apprenticeship Improving Science Education (iRAISE)*. Empirical Education Inc.
- Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M. A. (2019). A framework for learning from null results. *Educational Researcher*, 48(9), 580–589. <https://doi.org/10.3102/0013189X19891955>
- Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). *Impact of the Developing Mathematical Ideas professional development program on grade 4 students' and teachers' understanding of fractions* (REL 2017-256). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.
- Jerrim, J., Austerberry, H., Crisan, C., Ingold, A., Morgan, C., Pratt, D., Smith, C., & Wiggins, M. (2015). *Mathematics mastery: Secondary evaluation report*. Education Endowment Foundation.

- Kim, J. S. (2019). Making every student count: Learning from replication failure to improve intervention research. *Educational Researcher*, 48(9), 599–607. <https://doi.org/10.3102/0013189X19891428>
- Kuijpers, C. C. H., Fronczek, J., van de Goot, F. R. W., Niessen, H. W. M., van Diest, P. J., & Jiwa, M. (2014). The value of autopsies in an era of high-tech medicine: Discrepant findings persist. *Journal of Clinical Pathology*, 67(6), 512–519. <https://doi.org/10.1136/jclinpath-2013-202122>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, 24(1), 37–62. <https://doi.org/10.3102/01623737024001037>
- Lavrakas, P. J., Traugott, M. W., Kennedy, C., Holbrook, A. L., de Leeuw, E. D., & West, B. T. (Eds.). (2019). *Experimental methods in survey research: Techniques that combine random sampling with random assignment*. John Wiley.
- Levin, J., & Quinn, M. (2003). *Missing opportunities: How we keep high-quality teachers out of urban classrooms*. The New Teacher Project.
- Lindsay, J., Davis, E., Stephan, J., & Proger, A. (2017). *Impacts of Ramp-Up to Readiness after one year of implementation* (REL 2017-241). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest.
- Louisiana Believes. (2019a). *Instructional materials evaluation—Student standards review: enVision Math Common Core*. https://www.louisianabelieves.com/docs/default-source/curricular-resources/pearson-envision-math-grades-k-6.pdf?sfvrsn=ac0a831f_10
- Louisiana Believes. (2019b). *Guidance for instructional materials review*. <https://www.louisianabelieves.com/docs/default-source/curricular-resources/guidance-for-textbooks-and-instructional-materials-reviews.pdf?sfvrsn=2>
- Maas, T., & Lake, R. (2015, January). *A blueprint for effective and adaptable school district procurement*. Center for Reinventing Public Education. <https://files.eric.ed.gov/fulltext/ED558568.pdf>
- Martin, J., McBride, T., Brims, L., Doubell, L., Pote, I., & Clarke, A. (2018, February 22). *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them*. <https://www.eif.org.uk/resource/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them>
- Maynard, R. (2006). Presidential address: Evidence-based decision making: What will it take for the decision makers to care? *Journal of Policy Analysis and Management*, 25(2), 249–265. <https://doi.org/10.1002/pam.20169>
- McLaughlin, M. W., & Mitra, D. (2001). Theory-based change and change-based theory: Going deeper, going broader. *Journal of Educational Change*, 2(4), 301–323. <https://doi.org/10.1023/A:1014616908334>
- Ming, N. C., & Goldenberg, L. B. (2021). Research worth using: (Re)framing research evidence quality for educational policymaking and practice. *Review of Research in Education*, 45(1), 129–169. <https://doi.org/10.3102/0091732X21990620>
- Modarres, M. (1993). *What every engineer should know about reliability and risk analysis*. Center for Reliability Engineering.
- Murphy, R., Weinhardt, F., & Wyness, G. (2018). *Who teaches the teachers? An RCT of peer-to-peer observation and feedback in 181 schools* (CEP Discussion Paper No. 1565). Centre for Economic Performance.
- National Implementation Research Network. (2016). *Active implementation practice and science*. <https://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-Briefs-1-ActiveImplementationPracticeAndScience-10-05-2016.pdf>

- Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (Eds.). (2015). *Handbook of practical program evaluation* (4th ed.). Jossey Bass/Wiley.
- Pane, J. F., McCaffrey, D. F., Slaughter, M. E., Steele, J. L., & Ikemoto, G. S. (2010). An experiment to evaluate the efficacy of cognitive tutor geometry. *Journal of Research on Educational Effectiveness*, 3(3), 254–281. <https://doi.org/10.1080/19345741003681189>
- Papay, J. P., Bacher-Hicks, A., Page, L. C., & Marinell, W. H. (2017). The challenge of teacher retention in urban schools: Evidence of variation from a cross-site analysis. *Educational Researcher*, 46(8), 434–448. <https://doi.org/10.3102/0013189X17735812>
- Papay, J. P., & Kraft, M. A. (2016). The productivity costs of inefficient hiring practices: Evidence from late teacher hiring. *Journal of Policy Analysis and Management*, 35(4), 791–817.
- Petroski, H. (1992). *To engineer is human*. Vintage Books.
- Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, 42(8), 424–432. <http://dx.doi.org/10.3102/0013189X13507104>
- Rayes, F., Oh, J., Lee, S. S., & Boruch, R. (2016). *Ambient positional instability among teachers in Minnesota public schools: 2010–2011 to 2014–2015*. http://repository.upenn.edu/gse_pubs/400
- Resendez, M., Azin, M., & Strobel, A. (2009). *A study on the effects of Pearson's 2009 enVision-MATH program: Final summative report*. PRES Associates.
- Rickles, J., Zeiser, K., & West, B. (2018). Accounting for student attrition in power calculations: Benchmarks and guidance. *Journal of Research on Educational Effectiveness*, 11(4), 622–644. <https://doi.org/10.1080/19345747.2018.1502384>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Roy, P., Styles, B., Walker, M., Morrison, J., Nelson, J., & Kettlewell, K. (2018). *Best practice in grouping student intervention A: Best practice in setting. Evaluation report and Executive Summary*. Education Endowment Foundation.
- Schweppenstedde, D., & Reid, A. (2015, September 18). *How failure can feed success: Using evidence on "what does not work" to improve services and external recognition* [Blog post]. <https://www.rand.org/blog/2015/09/how-failure-can-feed-success-using-evidence-on-what.html>
- Shojania, K. G., & Burton, E. C. (2008). The vanishing nonforensic autopsy. *New England Journal of Medicine*, 358, 873–875. <https://doi.org/10.1056/NEJMp0707996>
- Snook, I., Marshall, J. M., & Newman, R. M. (2003, January). *Physics of failure as an integrated part of design for reliability*. Paper presented at the IEEE Proceedings Annual Reliability and Maintainability Symposium; Tampa, FL. <https://doi.org/10.1109/RAMS.2003.1181901>
- Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education. *Educational Evaluation and Policy Analysis*, 32(3), 351–371. <https://doi.org/10.3102/0162373710373389>
- Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *Journal of Experimental Education*, 82(3), 334–357. <https://doi.org/10.1080/00220973.2013.813364>
- 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. <https://doi.org/10.3102/0162373709339524>
- Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works,

- for whom and under what conditions. *Educational Evaluation and Policy Analysis*, 42(3), 354–374. <https://doi.org/10.3102/0162373720929018>
- Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Robyn, A., Baird, M. D., Gutierrez, I. A., Peet, E. D., Brodziak de los Reyes, L., Fronberg, K., Weinberger, G., Hunter, G. P., & Chambers, J. (2019). *Intensive partnerships for effective teaching enhanced how teachers are evaluated but had little effect on student outcomes*. RAND Corporation. https://www.rand.org/pubs/research_briefs/RB10009-1.html
- Strauss, V. (2018, June 29). Bill Gates spent hundreds of millions of dollars to improve teaching. New report says it was a bust. *The Washington Post*. <https://beta.washingtonpost.com/news/answer-sheet/wp/2018/06/29/bill-gates-spent-hundreds-of-millions-of-dollars-to-improve-teaching-new-report-says-it-was-a-bust/>
- Styles, B., Clarkson, R., & Fowler, K. (2014). *Chatterbooks: Evaluation report and executive summary*. Education Endowment Foundation.
- Taylor, J. A., & West, B. (2020). Estimating teacher attrition for impact study design. *Educational Researcher*, 49(1), 68–70. <https://doi.org/10.3102/0013189X19880550>
- U.S. Chamber of Commerce. (2007). *Leaders and laggards: A state-by-state report card on educational effectiveness*. <https://www.uschamberfoundation.org/leaders-and-laggards/app/docs/2007.pdf>
- USDOE Office of Inspector General. (2018a, November 27). *Calculating and reporting graduation rates in Utah* (ED-OIG/A06R0004). <https://www2.ed.gov/about/offices/list/oig/auditreports/fy2019/a06r0004.pdf>
- USDOE Office of Inspector General. (2018b, January 11). *Calculating and reporting graduation rates in California* (ED-OIG/A02Q0005). <https://www2.ed.gov/about/offices/list/oig/auditreports/fy2018/a02q0005.pdf>
- Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Sullivan, K., Ruiz de Castilla, V., Fleming, D. R., Henry, C., Long, T., & Hughes Jones, D. (2018). Findings from a multi-year scale-up effectiveness trial of Open Court Reading. *Journal of Research on Educational Effectiveness*, 11(1), 109–132. <https://doi.org/10.1080/19345747.2017.1342886>
- Valentine, J. (2019). Expecting and learning from null results. *Educational Researcher*, 48(9), 611–613. <https://doi.org/10.3102/0013189X19891440>
- Weiss, C. (2002). What to do until the random assigner comes. In: F. Mosteller, & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 198–233). Brookings Institution.
- Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, 1997(76), 41–55. <https://doi.org/10.1002/ev.1086>
- Welsh, R. (2021). Assessing the quality of education research through its relevance to practice: An integrative review of research-practice partnerships. *Review of Research in Education*, 45(1), 170–194. <https://doi.org/10.3102/0091732X20985082>
- West, M., Ainscow, M., Wigglesworth, M., & Troncoso, P. (2017). *Challenge the gap: Evaluation report and executive summary*. Education Endowment Foundation.
- What Works Clearinghouse. (2015). *WWC Standards Brief: Attrition standard*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_attrition_080715.pdf
- Wigglesworth, M., Squires, G., Birchinal, L., Kalamouka, A., Lendrum, A., Black, L., Troncoso, P., Santos, J., Ashworth, E., & Britteon, P. (2018). *Friends for life: Evaluation report and executive summary*. Education Endowment Foundation.
- Wiggins, M., Parrao, C. G., Austerberry, H., & Ingold, A. (2017). *Foreign language learning in primary school*. Education Endowment Foundation.
- Wiggins, M., Sawtell, M., & Jerrim, J. (2017). *Learner response system evaluation*. Education Endowment Foundation.
- Wilkins, C., Gersten, R., Decker, L. E., Grunden, L., Brasiel, S., Brunnert, K., & Jayanthi, M. (2012). *Does a summer reading program based on Lexiles affect reading comprehension?*

- (NCEE 2012-4006). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Williams, K. R., & Mattson, S. A. (2006). Qualitative lessons from a community-based violence prevention project with null findings. *New Directions for Evaluation, 2006*(110), 5–17. <https://doi.org/10.1002/ev.183>
- Worth, J., Sizmur, J., Walker, M., Bradshaw, S., & Styles, B. (2017). *Teacher observation: Evaluation report and executive summary*. Education Endowment Foundation.
- Ye, T., Frisone, M., Hooks, T., & Boruch, R. (2016). *Ambient positional instability in New Jersey public schools: 1996–1997 to 2011–2012*. https://repository.upenn.edu/cgi/view-content.cgi?article=1401&context=gse_pubs
- Younie, S. (2006). Implementing government policy on ICT in education: Lessons learnt. *Education Information Technology, 11*, 385–400. <https://doi.org/10.1007/s10639-006-9017-1>
- Zinth, K. (2005). *State textbook adoption*. Education Commission of the States.