AERA Open
January-December 2023, Vol. 9, No. 1, pp. 1–13
DOI: 10.1177/23328584231171536
Article reuse guidelines: sagepub.com/journals-permissions

Designing Field Experiments to Integrate Research on Costs

A. Brooks Bowden
University of Pennsylvania

Although experimental evaluations have been labeled the "gold standard" of evidence for policy (U.S. Department of Education, 2003), evaluations without an analysis of costs are not sufficient for policymaking (Monk, 1995; Ross et al., 2007). Funding organizations now require cost-effectiveness data in most evaluations of effects. Yet, there is little guidance on how to integrate research on costs into efficacy or effectiveness evaluations. As a result, research proposals and papers are disjointed in the treatment of costs, implementation, and effects, and studies often miss opportunities to integrate what is learned from the cost component into what is learned about effectiveness. To address this issue, this paper uses common evaluation frameworks to provide guidance for integrating research on costs into the design of field experiments building on the ingredients method (Levin et al., 2018). The goal is to improve study design, resulting in more cohesive, efficient, and higher-quality evaluations.

Keywords: *cost-effectiveness, economic evaluation, economics of education, educational policy, experimental design, experimental research, policy evaluation design, program evaluation, randomized control trial, research methodology*

Introduction

Initiatives during the George W. Bush administration and the Obama administration set the stage for a "Golden Age of evidence-based policy" (Haskins, 2015). Together, these administrations stressed the importance of internally valid effectiveness evidence to inform decisions to more efficiently allocate public resources in education. In 2002, the U.S. Department of Education's Institute of Education Sciences (IES) was built upon this idea and set out to grow the experimental evidence available within the field of education. Randomized field trials were the focus of IES initially, as this method is the gold standard for effectiveness evidence (U.S. Department of Education, 2003). Such studies, if executed properly, tell us if a particular program (broadly defined) achieves the goals it was designed to improve.

Knowing if an approach was found to impact the outcome as intended is necessary but not sufficient for policymaking when deciding among alternative options (Harris, 2008; Monk, 1995; Ross et al., 2007). Effects must be considered alongside the costs required to produce them to maximize public investments (Levin, 2001; Levin & Belfield, 2015). Such analyses, broadly called economic evaluation, include cost-effectiveness and benefit-cost analysis (Levin et al., 2018). By including an economic component, the goal is to provide evidence of effectiveness and corresponding costs and implementation fidelity so that the evidence is useful in policy considerations of how to best allocate scarce resources and to improve the likelihood of successful replication (Belfield & Bowden, 2019).

Major funders of policy and program evaluation in the United States—including IES, the Department of Labor, USAID, and others—now require evaluations to include a cost component through a cost-effectiveness or benefit-cost framework to ensure that the change in resources that produced effects are published in addition to program impacts. This expansion of expectations among funders has been a major challenge for researchers. Importantly, coursework on the methods of economic evaluation and on conducting research to estimate costs within field trials has been largely absent from predoctoral training (Clune, 2002; Harris, 2008; Levin, 2001, 2013; Rice, 1997). Unfortunately, it is common to see cost-effectiveness discussions in studies be treated as "back of the envelope" calculations or reference a purchase price even when the effects were estimated with careful attention to assumptions and causality.

This paper addresses the growing demand for rigorous cost-effectiveness analysis by providing guidance on how to design a cost study within the context of randomized field trials to assess the resources used in improving educational outcomes. Broadly, the method presented here—*the ingredients method*—is not new (the primary text on the subject has three editions, with the most recent being Levin et al., 2018). This paper contributes a study design guide to integrate research on costs into randomized field trials.

The ingredients method was developed to provide a straightforward approach to conducting economic analyses in education and other public sectors (Levin, 1975, 2001, 2013). Although this method is widely accepted as a rigorous approach to evaluating costs (Cost Analysis Standards Project, 2021), there are misconceptions in the field regarding the data required for this research within larger evaluations. Common misinterpretations are that every resource, regardless of relation to the outcome, must be observed and that most data are collected via intensive interviews. These false perspectives point to the need for specific guidance concerning the data collection needed to apply the ingredients method to experimental designs and how including these approaches can deepen what is learned from effectiveness studies. To date, there is no methodological guidance available to researchers that directly informs design practices to collect the data needed to conduct high-quality and efficient research on costs within field experiments.

This paper serves as a "how-to guide" for designing research on costs that are intended to supplement foundational work by Levin on cost-effectiveness and on field trial design strategies to estimate treatment effects, explore impact heterogeneity, and examine implementation/fidelity (for example, see Boruch, 1997; Orr 1999; Rossi et al., 2003). This work focuses on cost-effectiveness and field experiments and can be used as a complement to other available guidance on the treatment of ingredients in estimating costs for specific types of programs (see, for example, Jones et al., 2019, on early childhood education costs; Bowden et al., 2017, on estimating costs of service mediation interventions; Bradshaw et al., 2020, on estimating costs of disciplinary programming). My intention is to provide a road map for researchers to better design education research by integrating data collection efforts on ingredients (resources or inputs) with the study's purpose and plan for estimating and understanding effects. Throughout the paper, I discuss real educational interventions to deepen the guidance and implications through concrete examples.

To begin, I analyze education research proposals to demonstrate the need for guidance and to identify aspects of study design where simple modifications could ease the integration of cost-effectiveness into evaluations. Then, I provide a brief review of the ingredients method. The remainder of the paper focuses on guiding the design of evaluations to ensure that costs and effects correspond and to use the conceptualized contrast between treatment and control conditions to design the data collection plan. Although my focus here is on education, the concepts and guidance are broadly applicable to policy and program evaluations within the social sciences.

State of Education Research Design

Given that the goal of this work is to provide researchers with useful guidance that is rooted in the design of education research, I examined proposals to conduct field experiments for the inclusion of cost research and the quality of the cost research proposed. I obtained data on funded IES Goal 3 Efficacy proposals from the National Center for Education Research from fiscal years 2014, 2015, and 2016 under the Freedom of Information Act. The sample includes 59 efficacy proposals with distribution by year of 18 proposals in 2014, 19 in 2015, and 22 in 2016. Three proposals were not available due to redactions and are not included in the sample.

Although these proposals are not reflective of recently funded work, this time period is significant because 2014 was the first year that the request for applications from IES encouraged the inclusion of a cost study within an efficacy trial, and by 2016 efficacy proposals were required to include a cost component. Even though several years have passed, some of these trials are ongoing due to COVID-related delays or through longitudinal follow-up research. I briefly discuss this research to motivate the need for the tools presented in this article.

I coded each proposal for an array of information including program characteristics, methods to estimate and understand impacts, and methods to evaluate the costs of the program. I also evaluated the integration of the economic or cost component into the "story" of the evaluation and the purpose and justification of the work. A justification that integrates a resource or cost-effectiveness theme could be a rationale related to scarce resources, tradeoffs, labor market ramifications, returns to educational investments, improving access to economic sustainability, or the importance of understanding costs when deciding among educational programmatic alternatives. A proposal that tells a cohesive story is important as a signal that the cost component is being efficiently integrated into the evaluation. A cohesive proposal is also important for the overall strength of the proposal because it demonstrates that each aspect of the proposed work reflects the theory of change and is intended to add to what is understood about the effectiveness of an approach.

My findings, summarized in Table 1, indicate that proposals would be strengthened by more thoroughly integrating research on costs in multiple sections, including the problem/importance section that justifies the work, research questions, and the analytic plan to collect and analyze data. This constitutes a major shift from existing practice where cost-effectiveness is often treated as an afterthought or a procedure that is akin to a budget audit. By carefully considering the purpose of the evaluation—including economic justifications—the study can be designed with clear alignment between the purpose or problem being addressed, the policy at hand, the outcomes being measured, and the questions being asked. Just as proposals provide details on the data collection plan and timeline, descriptions and plans should also address how data on the program's ingredients or resources will be identified and collected, and how the evaluation and cost-effectiveness component will address

TABLE 1
Summary of Proposal Data Analysis Demonstrating the Need for Stronger Integration of Cost Research Into Design

	2014	2015	2016
Total proposals in sample	18	19	22
Problem statements include economic justification	5	16	20
Research question(s) on costs listed	2	1	5
Analysis plan section on economic evaluation or costs	4	17	22
Method to estimate costs identified	4	6	7

Notes: Counts shown. In total, 59 proposals were reviewed. Due to nonresponse, sample data are missing three proposals in 2015 and one proposal in 2016.

the contrast between the treatment and control groups. As such, it is important to establish a research question for this aspect of the evaluation to guide this work and to allow for transparency in reviewing what was planned versus what was conducted and reported.

The following guidance provides a road map to integrate cost-effectiveness into the story of the proposal so that the resulting research may be better equipped to produce effects that are explained by complementary research on both costs and implementation. Recent standards of practice in economic evaluation also point to the importance of integrating economic research into the design of efficacy trials to ensure that the costs estimated correspond to the effectiveness estimate (Cost Analysis Standards Project, 2021).

Defining Cost-Effectiveness

The concept of costs is best conceived as the value of resources delivered to achieve an effect (Levin et al., 2018). The economic principle of opportunity cost underpins this conceptualization, as all resources, regardless of who provides or funds them, have value and cannot be used for other purposes while being used for a given educational approach. The resulting analysis provides the social cost of an approach, preferably with a supplemental analysis of how the costs were financed. The ingredients method offers a clear approach to estimating costs and conducting cost-effectiveness analysis (Levin, 1975). The method includes three main steps to estimate costs: (1) identify, describe, and quantify ingredients; (2) match ingredients to standardized price values; and (3) calculate costs; and in the case of cost-effectiveness, a fourth step would be to combine costs with effects into a cost-effectiveness ratio (for a thorough treatment of the ingredients method and cost-effectiveness, please see Levin et al., 2018).

In a cost-effectiveness analysis, this fourth step includes estimating cost-effectiveness ratios and comparing alternative policy options on the cost to produce an additional unit of a shared outcome. As mentioned previously, to calculate a ratio of the cost of an approach relative to effects, the costs and effects must correspond by reflecting the resources received by the treatment in contrast to the resources received by the control. Thus, cost-effectiveness ratios naturally reflect the evaluations that estimated them. As a result, there is some complexity in comparing metrics across studies (Levin & Belfield, 2015).

Cost-effectiveness is useful in sectors, like education, where early outcomes are valued that are not easily monetized for benefit-cost analysis. Early literacy development and attendance are two of many examples where the field agrees on the importance of these outcomes for children and outcomes where additional assumptions about their long-term relationship to income and adult health may not be relevant for decisions made at the state and local levels. Cost-effectiveness ratios provide a metric to compare alternative strategies to improve an outcome by relative costs to achieve that outcome. This metric is distinct from a benefit-cost ratio, where impacts are translated into long-term monetary economic benefits (dollars). Cost-effectiveness is intended to support decisions in a format that relies on fewer assumptions and is thus easily interpretable and relevant for policy.

All four steps are important in evaluating cost-effectiveness, but this paper focuses on Step 1 because this process of identifying, describing, and quantifying ingredients is the most salient aspect for designing field trials. If study teams adequately and efficiently collect data on the ingredients allocated to produce an impact, cost estimation and other potential options for analyses can be explored depending on the purpose and audience of the evaluation. However, without adequate data on costs, the ability of the evaluation to address a range of questions will be limited.

Designing for Correspondence Between Costs and Effects

When estimating costs to combine with effects from a field trial, the cost estimate quantitatively represents the change in resources that resulted in a change in outcomes. To accomplish this, there are three important considerations for cost-effectiveness when designing a field experiment: implementation, treatment contrast, and expected heterogeneity of effects. The cost estimate must reflect these three aspects of delivering an intervention and the effects, especially if the broader goal is to compare cost-effectiveness ratios of alternative policy options, broadly known as cost-effectiveness

TABLE 2
Ingredients Worksheet to Design Data Collection

Ingredients	Description	Questions	Source	Units
Comprehensive list, regardless of who pays	Known drivers of effectiveness and costs	Mirror description; guide data collection during program delivery	Identify instruments planned to integrate cost data collection or make a plan to collect cost data separately	Expected units; use in surveys
Personnel				
e.g., staff, teachers, coaches, counselors, volunteers, caregivers	e.g., education, training, experience, time, roles, etc.	e.g., education level, prior relevant training, time, duties	e.g., surveys, interviews, system data, program records, time logs	e.g., FTE, days, hours
Facilities				
Space for delivering programming; e.g., office, classroom, cafeteria	e.g., required size and furnishings, time	e.g., space size, extra plugs, technology, other accommodations, rented or reallocated, dedicated to the program only or shared	e.g., surveys, observations, interviews	e.g., square feet, rooms, etc.
Materials				
e.g., books, computers, workbooks, manipulatives	e.g., laptop or tablet, kit of books for the classroom	e.g., dedicated or shared, replacement timeline	e.g., records, known or average "life" of material, survey	e.g., kits, hours, laptops

Notes: Adapted from Levin et al. (2018). The categories are intended to guide practice and may need to be adapted. For example, categories for training and other inputs—such as rewards, food, and internet—may be relevant.

analysis (Levin 1975; Levin et al., 2018). These concepts are critical not only to ensure the precision of the match between costs and effects but also in the efficiency of the evaluation itself. This section proceeds by providing more information on how these common evaluation concepts relate to cost research and study design.

Implementation of the Treatment

To ensure that the cost estimate corresponds to the effectiveness estimate, the costs should be based on what was provided during the trial. This is important because it relates to the very mission of the cost-effectiveness framework—to examine a program's effects relative to its costs and to inform future implementation. Actual program delivery, and the contrast between treatment and control conditions, may be quite different from the program's design or how the program was previously delivered and evaluated. For costs to correspond to effects, the data on ingredients—descriptive, qualitative data and quantities—must be collected during program delivery to reflect implementation.

The first step in designing data collection strategies for ingredients is to outline the ingredients of the intervention and the data needed to describe, quantify, and estimate costs. In Table 2, the common template for the ingredients method is adapted for the design phase to support planning for data collection and to encourage the integration of costs and implementation data collection activities when possible.

As an example, let us focus on the recent efficacy trial of Zoology One (Gray et al., 2022). This study examined the effects, implementation, and costs of a kindergarten literacy curriculum (Zoology One, now called ARC Core). The initial step to design the study to estimate costs that correspond to effects was to review the program's goals, components, and theory of change to identify ingredients. Table 3 shows this initial list with some descriptions of each ingredient. Each ingredient also has a quantity and unit of measurement that is based on the design of the curriculum and that the study planned to deliver. During the study, the teacher's implementation of the curriculum was observed to determine if all of these components were delivered as planned or if practices and ingredients were adapted during the evaluation.

Ingredients data can be collected through a range of sources. Depending on the program, some data may exist in management information or extant databases. For all data that are not readily available, there are many options to gather ingredients data, such as observations, interviews, surveys, time logs, and focus groups. In the design template shown in Table 2, the column for data source serves as an early opportunity to begin considering options for ingredients data collection during the course of the study.

Importantly, the options or methods available to collect data on costs are often already employed in field experiments through the implementation study. When data are collected on costs and implementation through the same strategies, it is more efficient to combine efforts when possible, which would also reduce the burden of research on participants and reduce the costs of conducting evaluations. Table 4 provides examples of how data collection efforts on implementation can be expanded in simple ways to incorporate data on ingredients.

TABLE 3
Example of Initial Ingredients List for Classroom Curriculum

Ingredients	Description & Notes	Est. Quantity	Units
Personnel			
Kindergarten teacher	Time to deliver the curriculum. May need to include time to prep or to communicate with families given home component. Teacher experience may be important for variation.	120	Minutes/day
Principal	Time related to the curriculum and building community.	10	Minutes/week
Home reading	Books go home daily to be read by a parent/caregiver.	30	Minutes/day
Training			
Initial PD module	Day of PD before the school year. May require space.	1	Session
Coaching sessions	One visit per month by curriculum coaches.	10	Session
Materials			
Book collection	450+ books, leveled and integrating science content.	1	Collection
Assessment	Curriculum assessment to guide instruction	1	Assessment
curriculum materials	manuals for each unit of the curriculum, power word cards, etc.	1	Curriculum
Data Management			
Computer	Laptop in the classroom for the teachers to use.	1	Computer
Online software	Curriculum database and online support.	1	Subscription
Facilities			
Reading nook	Cozy area with books that encourage reading. May need to observe to list out ingredients involved.	1	Nook

Notes: Contents loosely based on the initial ingredients outlined for Gray et al. (2022). The ingredients were identified based on the curriculum's design, program materials, the theory of change, and existing practice. Quantity is estimated based on the planned study. This list is used in future steps to design data collection and to determine how much data are needed for the study.

Within the Zoology One literacy curriculum efficacy study example, the study included implementation research that aimed to examine fidelity among the treatment classrooms and the responses of teachers to this change in instructional approach. School-based data collection activities for this study were planned to include site visits, instructional observations, teacher and principal interviews, teacher surveys, and teacher time logs. The study team intentionally incorporated the cost component during the proposal planning phase so each of these data collection activities could have the dual purpose of also collecting data to estimate costs. Table 5 provides an example of how data collection of a set of ingredients was planned to make the study more efficient and to reduce the burden of data collection on schools and teachers.

In the design of the Zoology One study, the theory of change and extant literature indicated that teacher experience and qualifications may vary in ways that could be important for the productivity of the intervention. To be able to reflect this in the cost per student and to ensure the costs correspond to effects, these data were collected in the teacher survey. Classroom observations documented the time allocated to the literacy block, as well as documenting the overall resources available in the classroom. The study included a teacher time log at four points during the school year to observe how teachers allocated their time during and outside of school. Principals were interviewed about their time. Because of study limitations, it was not possible to collect data from families directly, but information about home reading was collected through the teacher survey.

Treatment Contrast

Another aspect of design that is critical to understanding the production of the impact is the contrast between what is received by the treatment group and the control group (Hamilton & Scrivener, 2018). In education, the control condition is often not a pure "control" condition as seen in laboratory studies but rather a condition where students continue along with activities as usual. Methods for the study of implementation focus on the relative difference between treatment and control conditions in field experiments both in theory and then in how that contrast is actualized (Century & Cassata, 2016). The theoretical or expected relative strength distinguishes between the treatment program as it is designed and the control condition as is it known from standard practice (often called business as usual [BAU]) or in theory (Cordray & Pion, 2006). Achieved relative strength then accounts for infidelity related to the theoretical relative strength for both treatment and control (Hulleman & Cordray, 2009). An evaluation should be designed with information about the expected or theoretical treatment contrast so that, during the trial, the achieved relative strength or actual treatment contrast is reflected in the resources received by each group.

TABLE 4
Integrating Data Collection on Costs and Implementation

Data Collection Approach	Ingredients Data
Observations	Observe the context and level of resource allocation as part of typical practice
	Observe treatment and control resources in use
	Identify unexpected resources
Surveys	Collect data from all schools/classrooms—treatment and control—on basic ingredients
Interviews/focus groups	Inquire about aspects of the program or business as usual (BAU) that require an iterative format
	Ask about supports that may have been missed, not easily observed (especially volunteer or home inputs), or indirect costs
	Invite feedback on the ingredients list and understanding of dosage/participation
	Can be a subsample so long as the sample is adequately represented in resource use
Time logs	Use if time allocated by teachers or other individuals (personnel) is difficult to quantify via survey
	Administer to treatment and control
	Multiple time points if possible

TABLE 5
Example of Data Sources for Ingredients

Ingredients	Questions/Content Needed	Sources
Personnel		
Kindergarten teacher	Experience and qualifications	Teacher Survey
	Time teaching	Observation
	Time to prep or other	Time log
Principal	Time related to curriculum and building community	Interview
Home reading	Books go home daily to be read by a parent/caregiver	Teacher survey

Notes: Contents are loosely based on the initial ingredients outlined by Gray et al. (2022). The ingredients worksheet was expanded to include the questions or content needed to describe and quantify each ingredient to estimate costs. The source column lists data collection activities that are planned to include data for both costs and implementation components of the evaluation.

Returning to our literacy curriculum example, the curriculum was designed to provide 120 minutes of literacy instruction. At the time of design, it was common in the district where the study took place for the literacy block to include 90 minutes of literacy instruction. The theorized contrast in instructional time between treatment and control is 30 minutes.

However, if during delivery the treatment classrooms average 110 minutes of instruction and the control classrooms average 100 minutes of instruction, the achieved relative strength in terms of minutes of instruction is actually 10 minutes of instruction. Observing treatment contrast is extremely important in cost-effectiveness because both effects and costs reflect the relative difference between the treatment and control conditions. Thus, it is important at the design stage to consider how large the contrast will likely be and where important variation might occur to plan observations and data collection accordingly.

The context in which the evaluation takes place and the resources that are delivered through BAU serve as a starting place for this aspect of the evaluation. If the context has highly prescriptive guidelines on the amounts of time and resources provided, determining what is changed by the treatment should be straightforward. However, if services and supports are varied, in both treatment and control conditions, where the contrast in services received is more muddled, additional focus and data are needed to specify contrast.

Additional complexity can arise through "service mediation interventions" where a treatment leads to changes in other services that contribute to the production of the outcome (Bowden et al., 2017). The mediator services may be included as indirect costs or cost savings, but importantly, the services occur prior to the measurement of the outcome of interest and relate to the outcome. From a replication perspective, the effect on the outcome would be influenced by the change in services that also mediate the outcome. By excluding indirect changes in the resources participants received, the ingredients and costs estimated will not reflect the value of all of the resources delivered to produce effects.

Relating this to the curriculum example, at the design stage the comparison was thought to be a simple replacement where the current literacy curriculum would be exchanged for the innovative curriculum being tested. However, as the treatment and control classrooms were

observed, the study team learned that the control classrooms used a range of literacy programming, meaning that the contrast was not a simple contrast. To capture the high variation in programming among the control condition classrooms, the study included detailed information in the teacher survey to observe what other literacy programs were being used. Information on the use of other programs (in addition to the curriculum offered as the treatment) was also important for the treatment condition, so the survey questions were administered to both groups. This understanding of treatment and control was key to the cost component of the study and to the interpretation of effects and variation in effects.

Heterogenous Effects

The production of impacts is highly influenced by the context, clients being served, and variation in treatment contrast among sites or classrooms resulting in heterogeneous treatment effects (for an excellent illustration, see Weiss et al., 2014, p. 782). In this scenario, the expected production of impacts and variation of those impacts is also useful when designing and integrating research on costs.

Although treatment effects may vary due to a range of reasons that are not planned, some interventions are designed to be tailored to the context to increase buy-in. Interventions may be more effective when there is local support and adaptation of an approach (McLaughlin, 1990, 1998). In these instances, the evaluation design can be planned to capture site-level or context-level variations in resources and support provided to both treatment and control conditions with a specific plan to identify variations in ingredients and resource use among sites (Bowden & Belfield, 2015).

In the curriculum example, the study relied upon a multisite design with two cohorts of students across two school years. As described previously, the cost study was designed in tandem with the effects and implementation components so that the data on costs could be used to unpack and explore variation in the contrast between conditions within the study sample, which would result in variation in treatment effects. The treatment contrast in the study differed across cohorts because the "business as usual" literacy blocks differed across the two cohorts of schools. One cohort was a wide range of programming with very little consistency and a large contrast with the treatment; the second cohort received a more consistent literacy block with guidance from the school district. With evidence to support an understanding of how the cohorts differed, the study team was able to return to the theory of change to undertake analyses to explore heterogeneity in effects.

Taxonomy of Educational Programs

Among the design considerations listed previously, treatment contrast is the most salient as it reflects the change in resources that result in the change in effects. Understanding this contrast allows us to estimate costs and effects more clearly, to better interpret our results, and to compare findings across evaluations of program alternatives. In this section, I provide a taxonomy of programs based on basic conceptualizations of treatment contrast among programs that are often studied in efficacy trials to support the design process for researchers. To confirm that the taxonomy applies to a wide range of educational interventions, during the review of 59 intervention efficacy trial proposals, all were examined against the classifications and all of the interventions easily fit within a category.

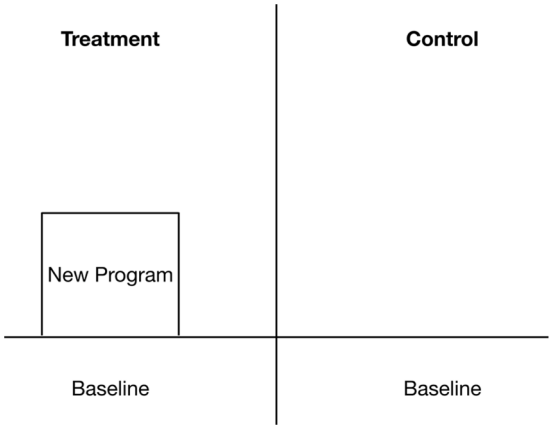

FIGURE 1. *Treatment contrast in resources for "new" programs.*

These classifications hinge on the treatment contrast—that is, the difference between what is delivered in the program being evaluated and what is delivered in the control or business-as-usual condition (Cordray & Pion, 2006; Hulleman & Cordray, 2009; Weiss et al., 2014). The classifications are *new*, *replacement*, and *supplemental*. A new program is an intervention that is unlike anything being provided and is in contrast to no service. A replacement program is intended to replace standard practice or business as usual. A supplemental or tiered program may involve an innovative approach that increases efficiency by partially replacing existing approaches or by adding onto existing programming to provide supplemental or tiered intervention support. In the following section, each type of program is described with examples and guidance for evaluation design. Importantly, these classifications are intentionally simple in nature as they are intended to serve as a guide and should be interpreted flexibly and adapted to each intervention and evaluation.

New Programs

The distinguishing factor for a new program is that it is an innovation that is being compared to no other similar

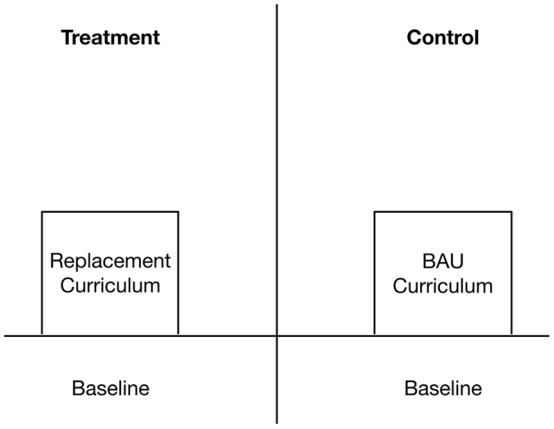

FIGURE 2. *Treatment contrast in resources for "replacement" programs, where the business as usual (BAU) curriculum currently used in practice is compared to a curriculum that replaces it.*

intervention. These interventions tend to occur outside of school time. For example, a volunteer tutoring program delivered to students during the summer is provided where the alternative to attending the program is no programming. As such, the evaluation is designed to randomly assign children to receive tutoring as the treatment compared to a control condition where students receive no similar services (summer, as usual). In this evaluation, the treatment is compared to the absence of similar programming or services, which means that the treatment contrast is equal to the full value of the treatment.

Figure 1 illustrates the contrast between treatment and control for an evaluation of a new program. Because there is nothing comparable being received by the control participants, the cost of the control condition is implicitly valued at zero.

To put this another way, this means that the "business as usual" (or, in the example, the summer as usual) inputs related to producing the outcome of interest are equivalent across the treatment and control groups such that the difference in costs between the groups is equal to the cost of the ingredients provided via the treatment intervention being evaluated (Cost of Treatment – Cost of Control = Cost of Treatment). In this case, where students in the control condition are not receiving services, there is no need to allocate scarce evaluation funding and time to collect data on the resources received through the control condition beyond what is needed to describe the control condition for replication purposes.

From a resource perspective, this type of program will typically include, at minimum, personnel, facilities, and materials. Although treatment contrast is most straightforward when evaluating a new program, heterogeneity among sites should still be considered. If the treatment's theory of change or implementation design is likely to result in variation in resources used to deliver the treatment, the data collected to estimate costs should reflect this variation and reflect important structural variation among sites.

Replacement Programs

A second type of program distinguished by its treatment contrast is a replacement program, where an innovative approach is replacing existing practice. In K–12, this type of program is likely occurring during school. For example, a curriculum or set of lessons is developed as an improvement on what is currently used in practice. In an evaluation, participants would be randomly assigned to receive the newly developed curriculum or the existing "business as usual" curriculum. In this instance, the contrast in resources used between treatment and control may be substantially overlapping because the existing approach and the replacement approach are delivered using equivalent time and space. Figure 2 illustrates this contrast (or lack thereof) showing an intervention that is designed to completely replace existing practice.

Here, the design implications are more complex than the new program contrast, where the difference in costs between treatment and control was simply equal to the cost of the treatment. When one practice or program replaces another, the production of the effect may be driven by how participants are taught or served, rather than through large changes in the resources provided. In this case, the study would need to be designed to capture any changes in resources to clearly describe what was provided through the treatment, what it was being compared to in the control, and what the difference between the two is (treatment – control) to estimate the cost of producing the effect clearly enough to support replication.

In schools, this is best described as a curriculum change where some resources are added (such as new textbooks, training, or materials) and some resources are changed (services received or supplemental support) but the base service of schooling goes unchanged. Although some studies assign the purchase price of a new textbook as the cost of the replacement curriculum (Chingos & Whitehurst, 2012; Boser et al., 2015; Kodel et al., 2017), this approach may overlook important costs that must be included when estimating the costs to produce effects. Overlooked costs may include expensive inputs, such as training and coaching, or inputs that are less expensive or not incurred by the school, such as volunteer time or homework. In addition, this approach overlooks teacher time, which is often invisible because so much of their work falls outside of the contracted 180 days of instruction (Hess, 2017).

Another important consideration about teacher time is that a teacher's bandwidth for reform is likely limited related to "programitis" (Elmore, 2004; Murnane & Nelson, 2007). Each curriculum comes with its own system for classifying

TABLE 6
Replacement Programming and Treatment Contrast Design Guide

Directions: Replacement programs will likely include all types of ingredients. The following questions are intended to simplify the data collection planned for personnel and serve as examples of how other inputs can be approached.
Will the main personnel delivering the program be randomly assigned to treatment and control groups? Or will the staff in these groups have similar characteristics?
Yes: Following equivalent groups, there is no need to collect information on qualifications.
No: Collect data on qualifications that are relevant for treatment effects and costs.
Will the main personnel in treatment and control conditions spend equal time delivering programming? (This also assumes no variation within each condition.)
Yes: Consider simple ways to confirm no differences across or within conditions.
No: Include time logs or other ways to measure time allocation over the course of delivery.
Will support staff be equally available across treatment and control conditions? (This also assumes support staff is not a key aspect of the theory of change.)
Yes: No need to collect information or consider confirming via survey or select interviews.
No: Collect data on support staff similar to main personnel.
Will treatment and control conditions use volunteers and caregivers equally? (This also assumes volunteers and caregivers are not a key aspect of the theory of change.)
Yes: No need to collect information or consider confirming via survey or select interviews.
No: Collect data on time and, if necessary, qualifications and travel; may be through main personnel if direct contact with volunteers and caregivers is not possible; consider limitations of data through sensitivity analyses.
If the treatment group providers receive training, is the training replacing current professional development? (This is asking about staff time, which is separate from external training costs or ingredients.)
Yes: Confirm time in training, including the need for substitutes, is equal across groups.
No: Collect data on the time spent in training, if a substitute was needed, etc.
Does the treatment require additional support or training for administrators?
Yes: Include the additional time; determine mode of data collection based on the amount of time and the role of administrative support in the theory of change (as time increases, amount of data increases).
No: No need to collect information.

levels of skill and providing differentiated instruction, books, and materials, and often includes differing approaches to teaching. The teacher in this scenario is not simply grabbing a new textbook off the shelf. Hidden activities may include attending coaching throughout the year, preparing for teaching with a new approach and/or materials, developing complementary materials, working with other teachers to build a learning community, and communicating with families.

Although replacement programs are more challenging to evaluate than a new program, data collection can be simplified by focusing on the resources being changed. This may mean that only a portion of the full resources (sometimes called incremental costs) need to be examined to estimate the cost to produce an effect—for example, an evaluation of an algebra curriculum that randomly assigns classrooms to receive the replacement curriculum or to continue with typical mathematics instruction. Both use the same facilities (classrooms) and time (core subject instruction time) to deliver the mathematics block. The goal of the evaluation is to determine if the replacement algebra curriculum does a better job of teaching math skills than current practice. In the context of the evaluation, schools exist and provide transportation and meals, and teachers and classrooms are in use following typical practice. Beyond describing typical resource allocation and the context for replication, the resources that are provided broadly as part of school are not required to estimate the cost of the curriculum change. Thus, the study should focus on aspects of schooling that are altered due to the curriculum change and relevant for the production of the outcome of interest.

Ingredients for replacement programs in schools will likely include teacher time, training, and textbook materials. Some programs will also include prizes, parent/caregiver time, additional instruction, additional instructional materials, and opportunities for experiential learning. If the program leads to the reallocation of other resources—such as supplemental staff, psychologists, vice principals, and special education supports—these indirect changes should be observed as they relate to the costs (or even cost savings) in producing the outcome (Bowden et al., 2017).

To support design, Table 6 lists questions to provide a starting point when planning to evaluate the cost of a replacement program within a field experiment. The questions suggested may need to be adapted because cost studies do not follow a one-size-fits-all approach. As described previously, the study is guided by the theory of change. One way to simplify the design phase is to focus on personnel.

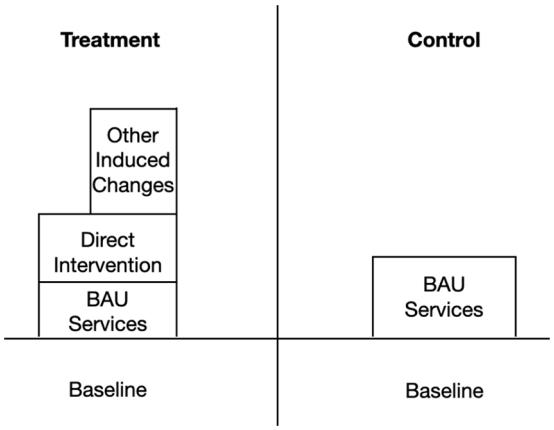

FIGURE 3. *Treatment contrast in resources for "supplemental" programs, where some services are provided as BAU to all students in potentially varying amounts.*

Personnel is an expensive, and likely the most important, input in education and other social services. Thus, the questions in Table 6 are focused on the people who deliver the program, but these questions can serve as a guide for other types of ingredients.

Supplemental or Tiered Support Programs

Supplemental programs or tiered support programs offer specialized services based on developmental needs related to the outcome of interest. These services are varied, often depending on the context, the type and scale of the support provided, and the needs being addressed. Although some aspects of supplemental programs may replace existing practices, these services are most easily defined as having a theory of change that is additive and complementary to typical business as usual practice. Another complexity is that many of these services are required by law, making a true control condition impossible and where most participants are receiving a range of services at any given time. Because of these two components, identifying the contrast in resources received between treatment and control conditions can be very challenging to plan for and to carry out during an evaluation.

In Figure 3, this complexity is illustrated by a control condition receiving more "BAU services" than the treatment condition. The treatment condition in the figure also received "BAU services" and the "direct intervention services" being evaluated as the intervention of interest. The figure also shows that other supplemental services "other induced services" are changed as a result of the treatment. Although the figure shows these induced changes as additive, please note that in some instances they may occur as cost savings.

In education, supplemental programs may focus on supporting academic skill development or on supporting students' comprehensive strengths and needs related to health, development, or other barriers to learning (see, for example, Bowden et al., 2020; Jacob et al., 2016). Law requires that students who do not have mastery of content and skills reflecting their grade level receive supplemental support in academics. Schools also often provide services or support for children's nonacademic challenges.

Because business as usual typically includes many supplemental support services, this category has the least clear distinction between what is received by each experimental group. The students in both groups may receive a range of services in addition to the treatment being evaluated. The services provided may reflect treatment assignment, where school services are reallocated to the children in the control condition to compensate for their group placement (or to use the resources that are now not being used by the treatment group). Further muddling the contrast between conditions, the baseline captured by "school" may also be altered through whole-school approaches to integrate support services, target discipline and behavior, and reduce chronic absenteeism.

In this case, it is critical to carefully consider resource implications of the theory of change and the data needed to understand the services and supports provided to treatment and control conditions, as well as other changes in services related to the outcome of interest. For example, in an evaluation of a whole school comprehensive student support model, schools are randomly assigned to receive the model of interest or to continue providing school as usual. The model aims to improve student engagement (attendance, GPA, and suspensions), achievement (ELA and math test scores), and attainment (on-time grade progression and graduation). In all schools in the study sample, typical "school as usual" includes efforts to address school climate and safety and social and emotional learning programing. These existing practices target the same outcomes as the whole school student support model being evaluated. To design the study, researchers should begin by examining information about existing programming to identify overlaps in theories of change across the intervention being evaluated and typical practices that are present in schools in the sample. In this case, resource, data on each approach that is theorized to impact the outcomes of interest would need to be collected within schools in both experimental conditions to ensure that costs correspond to effects.

Not all supplemental programs are this complex. If the treatment is directly linked to an outcome, the design is more straightforward. For example, a supplemental reading program is evaluated for effects on reading skills. Thus, the evaluation should be designed to examine the resources delivered through the treatment and the other types of services within the school that also target reading skill development.

Table 7 provides examples of questions that are important to consider during the design phase when evaluating a supplemental support program or tiered support program. The

TABLE 7
Supplemental Programming and Treatment Contrast Design Guide

Directions: Supplemental or tiered service programs will likely include all types of ingredients. The following questions focus on the overlapping nature of services and the challenge of measuring services across both experimental groups.
For school-based programs, does the intervention occur during school?
Yes: Determine if the intervention is changing how school is provided and/or if other aspects of schooling are being displaced by the intervention. For example, does the intervention require class time and staff meetings, or are other aspects of the day changed to accommodate the approach? Is space required for meetings or staff?
No: Consider staff time and facilities space. Determine if other external activities and services are being replaced. Transportation is likely necessary.
Based on the theory of change, does the program integrate aspects of other programming or services or does the program likely change other services received by participants?
Yes: Plan to observe changes in other services as a result of the program. For example, are other services better matched, required less, or used more often?
No: Focus on the theory of change and target data collection on the intervention. Check in at regular intervals to confirm no activities or other services were changed as a result of the program.
Does the program offer tiered support or individualized services?
Yes: Plan to collect data on resources related to how the intervention identifies who will receive which services; if staff monitor service participation and progress; and if caregivers and families are involved in the process and if they are providing time, transportation, or other inputs. Plan to collect ingredients, dosage, and intensity of services to estimate the cost of each service provided, as well as information about how many participate (if individual participation data are not available).
No: Focus on how many students are served, any variation in dosage, and if the program was not used to the scale at which it was designed. For example, if a program intends to serve 40 students but only 30 participated, it is important to note how the costs would change if all 40 students had participated.

questions are focused on the difficult nature of identifying clearly what is offered by the intervention that is distinct from what is received by both the treatment and control groups typically. A key identifier to guide the ingredients list for a supplemental program is the time at which the intervention occurs. If the program is before or after school, it is a signal to consider staff, space, transportation, and student time (for older students with foregone wages).

Although these questions are not exhaustive, they provide a head start on how to conceptualize the cost component of a study during the proposal and design phase to ensure that data collection activities incorporate the information needed to estimate costs accurately.

Returning to the Efficacy Trial Example

Briefly, consider the literacy curriculum study described in the prior section on designing for correspondence between costs and effects. To design the study, the curriculum would be considered a "replacement" program because the literacy curriculum replaces existing practice. As described previously, the students all go to school for the same amount of time so there is no need to consider transportation. The same logic applies to lunch and school facilities and furnishings, etc. The teachers were randomly assigned, and there was balance across groups in experience and training. At the time of study design, as mentioned previously, there was the expectation that instructional time would differ, so observations and teacher time logs captured these data (as described in Table 5). Questions about other instructional staff were included in interviews to confirm that there was no difference across groups. The home reading component was a large change from typical BAU practice in kindergarten, so home reading was included in the teacher survey to examine the costs borne by families related to the effects of the curriculum.

Although the curriculum as described is clearly categorized as a replacement, if it were offered over the summer in contrast to no literacy instruction, the contrast would be categorized as "new," reflecting the much larger contrast. In this case, the study would need to include facilities, staff time, transportation, and any other ingredients required to deliver the curriculum over the summer in addition to what was described previously.

Again, this taxonomy is not intended to be rigid, but it is intentionally simple. Some programs have a very clear treatment contrast that can be used during the design and proposal phase as a guide to focus scarce research resources. The goal is to collect data on the ingredients that matter most to ensure that estimated costs correspond to effects, which in turn supports the interpretation and comparison of the effects for policy and practice.

Conclusion

This paper aims to ease the burden of integrating research on the costs of field experiments by guiding how to design and propose evaluations based on the anticipated treatment contrast. This paper is also intended to improve the cohesion and efficiency of research by highlighting opportunities to integrate data collection on costs with implementation

research. The conceptual foundation for the classification system proposed here pushes researchers to use the theory of change to design a field trial around the theorized production of impacts and the treatment contrast. Broadly, programs fall into three categories—*new*, *replacement*, and *supplemental*—based on what would have occurred in the absence of the intervention and how much of business as usual is replaced or changed. Although these categories are not intended to be interpreted strictly, all the interventions described in the proposals in the sample reviewed in crafting this paper could be classified with this basic taxonomy. Future research should continue to build upon this system to develop additional design tools and aids to continue to improve the quality and efficiency of economic evaluation.

Acknowledgments

I would like to thank the following individuals for their feedback on this manuscript: Henry Levin, Steven Bell, Thomas Brock, Rebecca Maynard, Abigail Gray, AERA Open Editor Michal Kurlaender, and anonymous reviewers. Thank you to Rebecca Davis, Sangyoo Lee, and Viviana Rodriguez for research assistance and support. I appreciate receiving valuable input from the staff at the Institute of Education Sciences. I am also grateful to attendees of the annual meetings of the Association for Public Policy Analysis and Management; the Society for Research on Educational Effectiveness; the IES Methods Training Program in Economic Evaluation; and seminar attendees at the University of Pennsylvania, University of California Santa Barbara, New York University, and University of Virginia.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200034. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

JEL Codes

C18, C81, C93, D24, D61

ORCID iD


A. Brooks Bowden https://orcid.org/0000-0001-6079-5456

References

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Sage.

Boser, U., Chingos, M., & Straus, C. (2015). *The hidden value of curriculum reform: Do states and districts receive the most bang for their curriculum buck?* Center for American Progress.

Belfield, C., & Bowden, A. B. (2019). Using resource and cost considerations to support educational evaluation: Six domains. *Educational Researcher*, *48*(2), 120–127.

Bowden, A. B., & Belfield, C. R. (2015). Evaluating TRIO: A benefit-cost analysis and cost-effectiveness analysis of talent search. *Journal of Benefit-Cost Analysis*, *6*(3), 572–602.

Bowden, A. B., Shand, R., Belfield, C. R., Wang, A., & Levin, H. M. (2017). Evaluating educational interventions that induce service receipt: A case study application of city connects. *American Journal of Evaluation*, *38*(3), 405–419.

Bowden, A. B., Shand, R., Levin, H. M., Muroga, A., & Wang, A. (2020). An economic evaluation of the costs and benefits of providing comprehensive supports to students in elementary school. *Prevention Science*, *21*(8), 1126–1135.

Bradshaw, C. P., Debnam, K. J., Player, D., Bowden, B., & Lindstrom Johnson, S. (2020). A mixed methods approach for embedding cost analysis within fidelity assessment in school-based programs. *Behavioral Disorders*. https://doi.org/10.1177/0198742920944850

Century, J., & Cassata, A. (2016). Implementation research: Finding common ground on what, how, why, where, and who. *Review of Research in Education*, *40*(1), 169–215.

Chingos, M. M., & Whitehurst, G. J. (2012). *Choosing blindly: Instructional materials, teacher effectiveness, and the common core*. Brookings Institute, Brown Center on Education Policy.

Clune, W. H. (2002). Methodological strength and policy usefulness of cost-effectiveness research. In H. M. Levin, & P. J. McEwan (Eds.), *Cost-effectiveness and educational policy* (pp. 55–70). Eye on Education.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin, & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). American Psychological Association.

Cost Analysis Standards Project. (2021). *Standards for the economic evaluation of educational and social programs*. American Institutes for Research. https://www.air.org/sites/default/files/Standards-for-the-Economic-Evaluation-of-Educational-and-Social-Programs-CASP-May-2021.pdf

Elmore, R. (2004). *School Reform from the Inside Out: Policy, Practice, and Performance*. Cambridge, MA: Harvard Education Press.

Gray, A., Sirinides, P., Fink, R., & Bowden, A. B. (2022). Integrating literacy and science instruction in kindergarten: Results from the efficacy study of Zoology One. *Journal of Research on Educational Effectiveness*, *15*(1), 1–27. https://doi.org/10.1080/19345747.2021.1938313

Hamilton, G., & Scrivener, S. (2018). *Measuring treatment contrast in randomized controlled trials*. MDRC Working Paper.

Harris, D. N. (2008). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3–29.

Haskins, R. (2015, March 11). *Interview with Ron Haskins, APPAM President-elect*. http://www.appam.org/interview-with-ron-haskins-appam-president-elect/

Hess, F. (2017, September 11). Educator's time loss and the invisible cost of reform. *Education Week*. http://blogs.edweek.org/edweek/rick_hess_straight_up/2017/09/the_invisible_cost_of_reform.html

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88–110. http://dx.doi.org/10.1080/19345740802539325

Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness, 9*(Sup1), 67–92.

Jones, D. E., Bierman, K. L., Crowley, D. M., Welsh, J. A., & Gest, J. (2019). Important issues in estimating costs of early childhood educational interventions: An example from the REDI program. *Children and Youth Services Review*, *107*, 104498. https://doi.org/10.1016/j.childyouth.2019.104498

Kodel, C., Li, D., Polikoff, M. S., Hardaway, T., & Wrabel, S. (2017). Mathematics curriculum effects on student achievement in California. *AERA Open*, *3*(1), 1–22.

Levin, H. M. (1975). Cost-effectiveness analysis in evaluation research. In M. Guttentag, & E. L. Struening (Eds.), *Handbook of evaluation research* (*vol. 2*). Sage.

Levin, H. M. (2001). Waiting for Godot: Cost-effectiveness analysis in education. *New Directions for Evaluation*, *90*, 55–68.

Levin, H. M. (2013). Cost-effectiveness evaluation in education. In M. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 180–188). Sage.

Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*,*8*(3), 400–418. 10.1080/19345747.2014.915604

Levin, H. M., McEwan, P. J., Belfield, C. R., Bowden, A. B., & Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis* (3rd ed.). Sage.

McLaughlin, M. W. (1990). The Rand change agent study revisited: Macro perspectives and micro realities. *Educational Researcher*,*19*(9), 11–16. https://doi.org/10.3102/0013189X019009011

McLaughlin, M. W. (1998). Listening and learning from the field: Tales of policy implementation and situated practice. In: A. Hargreaves, A. Lieberman, M. Fullan, & Hopkins D. (Eds.), *International Handbook of Educational Change. Kluwer International Handbooks of Education* (*vol. 5*). Springer. https://doi.org/10.1007/978-94-011-4944-0_4

Monk, D. H. (1995). The costs of pupil performance assessment: A summary report. *Journal of Education Finance*, *20*, 363–371.

Murnane, R. J., & Nelson, R.R. (2007). Improving the performance of the education sector: The valuable, challenging, and limited role of random assignment evaluations. *Economics of Innovation and New Technology*, *16*(5), 307–322. https://doi.org/10.1080/10438590600982236

Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. Sage.

Rice, J. K. (1997). Cost analysis in education: Paradox and possibility. *Educational Evaluation and Policy Analysis*, *19*(4), 309–317.

Ross, J. A., Barkaoui, K., & Scott, G. (2007). Evaluations that consider the cost of educational programs: The contribution of high quality studies. *American Journal of Evaluation*, *28*(4), 477–492.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach* (7th ed.). Sage.

U.S. Department of Education (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Prepared by the Coalition for Evidence-Based Policy and published by the U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

Author

BROOKS BOWDEN is an assistant professor at the University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104; email: bbowden@upenn.edu. Her methodological research aims to improve the quality of education research with the goal of improving outcomes for children around the world.