Using Resource and Cost Considerations to Support Educational Evaluation: Six Domains

Clive R. Belfield[1] and A. Brooks Bowden[2]

Cost, cost-effectiveness, and benefit-cost analysis are methods used by economists to evaluate public policies. Essentially, these methods rely on impact evaluations, that is, research studies of efficacy and effectiveness. However, in most research in education, these cost and impact evaluations are performed separately. This separation creates methodological deficiencies and undermines the contribution of educational research to decision making. In this article, we identify key domains of educational research evaluations that, we believe, would be enhanced if resource and cost analyses were integrated more directly. These domains relate to outcome specification, treatment contrast, implementation fidelity, the role of mediators, power of the test, and meta-analysis. For each domain, we provide a case study example of how these cost analyses can complement and augment current research practices in educational evaluation. More interaction between economists and education researchers would be beneficial for both groups.

Keywords: econometric analysis; economics of education; program evaluation

Economic methods for program evaluation are cost analysis (CA), cost-effectiveness analysis (CEA), and benefit-cost analysis (BCA). These analytical approaches can be readily adapted to evaluate educational interventions and reforms with the goal of improved decision making (see Boardman, Greenberg, Vining, & Weimer, 2018; Levin, McEwan, Belfield, Bowden, & Shand, 2017). CA requires accurate estimation of all the extra resources required to implement a new intervention (e.g., more teachers, more classroom space); this tells the decision maker if the intervention is affordable. CEA compares these extra resources to gains in an educational outcome such as increased test scores or graduation rates; this tells the decision maker what this intervention costs per unit of educational improvement. BCA compares these extra resources against the monetary consequences of implementing an intervention, such as higher earnings from having more high school graduates; this tells the decision maker if the returns to the intervention exceed the costs of investment. In themselves, these three approaches—which we refer to here as cost analyses—are useful for evaluation.

In an essential way, however, cost analyses are linked to broader educational evaluations (on which, see Song & Herman, 2010; Weiss, Bloom, & Brock, 2014). To determine cost, it is first necessary to understand how an intervention is implemented. To determine cost-effectiveness, it is first necessary to identify effectiveness. Finally, to calculate money benefits, it is first necessary to know what impacts an education program has. Without this prior information from an educational research study, these cost analyses cannot proceed.

Our claim in this article is that too little attention has been paid to how these two types of evaluation cohere and especially to how consideration of resource use can guide impact evaluation. We illustrate how CA, CEA, and BCA can enhance the quality and utility of impact evaluations. That is, we imagine how educational research might be shaped if it was intended to be linked to some type of cost analysis.

We highlight six domains where a more extensive application of these three analytical approaches would be beneficial. Table 1 shows a summary of the domains, along with the relevant economic analysis and its implications. These domains are not a unified set: They are simply domains where consideration of economics can enhance the validity and utility of impact evaluations. The domains reflect where cost-based analysis may be the

[1]Queens College, City University of New York, New York, NY
[2]North Carolina State University, Raleigh, NC

Educational Researcher, Vol. XX No. X, pp. 1–8
DOI: 10.3102/0013189X18814447
Article reuse guidelines: sagepub.com/journals-permissions

Table 1
Education Research Method and Economic Method

Education Research Domain	Relevant Economic Method	Implications From Applying Economic Methods to Educational Impact Evaluations
Outcome specification	Benefit-cost analysis	Outcomes should not be overspecified, collinear, but may fade out
Treatment contrast	Opportunity cost	Counterfactual should be specified in as much detail as treatment
Service mediation	Indirect cost analysis	Resources that flow from treatments should be specified in detail and compared to size of treatment
Implementation fidelity	Direct cost analysis	Formal economic metrics to determine implementation fidelity should be considered
Power of the test	Cost-effectiveness analysis; benefit-cost analysis	Minimum Detectable Effect Size should be linked to Minimum Detectable Economic Consequence; power calculations may be revised
Meta-analysis	Cost analysis; cost-effectiveness analysis	Average effect size should be linked cautiously to expected cost; moderators should be checked as to how they relate to resource use

Note. See also Levin et al. (2017, Table 1).

most useful; each domain is separate from the others. Not every domain is relevant for every study, and it is unlikely all domains are relevant for any single study. But each domain covers an important part of the research process; as such, cost-based analyses are relevant across many research studies. As a priority, we recommend greater attention to how economic analyses can improve "outcome specification." The choice of outcomes is critical for all impact evaluations, and both CEA and BCA can make multifaceted contributions. A second priority is "service mediation"; it too is applicable across many educational interventions.

For each domain we provide examples of how, and instances where, cost analyses can complement and augment current research practices in education. These examples are illustrative but broadly representative of current research practice by education researchers (and by many economists researching in education). Our goal is not to single out individual studies for criticism but instead to promote evaluation methods that yield results which are more policy-relevant and improve decision-making.

Outcome Specification

The essential question for impact evaluation is, What are the outcomes of interest, that is, the potential outcomes that would occur in different contexts or circumstances (Weiss et al., 2014)? From an educational perspective, possible responses include measures of cognition, achievement or test scores, or years of attainment.

We argue for the choice of outcomes to be guided more explicitly by resource use considerations. That is, outcomes should be prioritized insofar as they affect how society uses its resources (Heckman & Kautz, 2012). This approach emphasizes changes in behavior over changes in students' achievement, test scores, or other measures of cognitive functioning (Duckworth & Yeager, 2015). Behavioral changes in turn connote resource use change; gains in cognition (however defined) or other psychological constructs may involve behavioral change but need not; they may simply change a student's knowledge level. The distinction is not semantic. The canonical example is the HighScope Perry Pre-school Program (Nores, Belfield, Barnett, & Schweinhart, 2006). This program exhibited complete fade-out in IQ gains by age 12 but did cause significant changes in behavior, including criminal activity and labor market participation. For evaluation purposes, these substantial behavioral consequences rendered the weak impacts on IQ less important (although still of academic interest).

Using an economic approach, many behavioral changes can be expressed in dollars and so compared to costs to perform a BCA (see Levin et al., 2017, Chapters 9–10). However, even absent BCA, a greater focus on behaviors yields several methodological contributions.

First, this focus should motivate evaluators to more clearly justify the use of test scores as outcome measures. Higher test scores may affect behavior, but the strength and consistency of the link—not just for earnings but also for health status and criminal activity—cannot be presumed. Another important study finds that long-term behavioral change is not mediated through test score changes. In their study of classroom quality, Chetty et al. (2011) identify gains in short-term achievement if classroom quality increases; as well, students in higher quality classrooms earn more in adulthood. Yet they find no medium-term gain in achievement. Whatever the mechanism is by which classroom quality affects labor market outcomes, it is not consistently evident in test scores. Chetty et al. offer several possible reconciliations (e.g., short-term achievement is a weak or unstable proxy). Each reconciliation implies that achievement is neither stable nor especially valid way to measure behavioral change.

Second, BCA may provide guidance on which set and how many outcomes to evaluate. In BCA, all benefits (monetized impacts) must be additive. This influences the choice and number of outcomes. It is invalid to add two benefits that represent the same behavioral change; only separable, independent benefits can be added up. For example, college might boost earnings and improve health; it is not valid to add in any health-induced earnings gains to the sum of benefits of college (although it is valid to include health gains that are not related to earnings).

Finally, BCA may help evaluators address the problem of "fade-out," that is, the likelihood that impacts decline to zero shortly after an intervention is terminated (Castleman, Page, & Schooley, 2014). Fade-out is common in education studies ("far too often" according to Bailey, Duncan, & Watts, 2017; see

Protzko, 2015, for an emphatic statement on cognitive fade-out). Hence, policy makers might reasonably ask why an effective educational program should be implemented: Sooner or later, they might argue, test scores will revert back to the average. One response is to refer to behavioral changes and whether the discounted benefits exceed the costs before fade-out happens. An intervention may still be valuable even with rapid fade-out. We can imagine an intervention that boosts students' reading in third grade but where scores revert to the mean by fifth grade. We are not sure what an impact evaluation might recommend in this scenario. However, the intervention might be justifiable if fewer reading counsellors are needed for two grade years—that would change resource use. As fade-out happens "far too often," it becomes especially important to focus on resource use.

We illustrate the influence of BCA on outcome specification with evidence on socioemotional learning (SEL). Socioemotional skills, which include competencies such as self-awareness, self-management, and responsible decision making, are recognized as important for child development. In reviews of over 200 studies, Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011, Table 2) and Sklad, Diekstra, De Ritter, Jehonathan, and Gravestein (2012) identify durable and substantively important gains in five outcomes: (a) achievement, (b) attitudes, (c) social behavior, (d) conduct, and (e) emotional states. *Prima facie*, these gains make a persuasive case for policies to promote SEL.

These outcomes should be assessed in several ways (for a full discussion, see Belfield et al., 2015). First, it is important (or, we might argue, necessary) to establish that each outcome does connote a change in behavior and thereby a change in resource use. Changes in academic performance (a) might not influence behavior; similarly, changes in attitudes (b) should be validated against changes in behavior. (We are not denigrating attitudinal studies; attitudinal change might be powerful, but it needs to be justified.) Also, deviant social behaviors (c) may be more deleterious than deviant conduct (d); a useful analogy is the criminology distinction between felonies and misdemeanors. Second, the magnitude of behavioral change needs to be identified so it can be aggregated. The changes across the five realms might affect lots of behaviors, both for individual students and their peers; the effects may be long-lasting or temporary; they may appear immediately or after a lag. Finally, all potential outcomes must be independent. Across the five realms listed above, some outcomes may be independent; others are likely to be collinear (e.g., achievement with attitude or conduct problems with emotional distress). It may be misleading to aggregate gains reported in the review literature. When there are multiple reported outcomes, it is essential to establish their independence.

Overall, BCA offers guidance about the types, number, and attributes of potential outcomes for evaluative research in education. (We focus on student-level outcomes, but our argument extends to teacher-, classroom-, or school-level outcomes.) Specifically, measured outcomes should be based on metrics that have direct behavioral or resource consequences; they should cover the duration of the impact of the intervention; and the extent of their independence from each other should be established.

Treatment Contrast

For evaluation of any educational intervention to be valid there must be a treatment contrast: "If a treatment contrast does not exist ... there cannot be a program effect. Thus, a treatment contrast is necessary for a program effect to occur" (Weiss et al., 2014, p. 785). Although evaluators typically do specify the treatment in detail, much less attention is paid to the counterfactual experience and hence to the actual treatment contrast. A greater emphasis on economic evaluation should help redress this imbalance.

A key principle of economics is opportunity cost—that is, policies should be valued in terms of the best alternative use of resources. When evaluating, say, *Reading Recovery*, the opportunity cost is what the reading teachers would have otherwise done (or indeed if math teachers could have been hired instead). This principle embodies the idea of treatment contrast: the appropriate way to value an intervention is to look at what is given up by implementing that intervention. An economic evaluation therefore forces (or should force) the researcher to describe the counterfactual in as much detail as the treatment. By identifying what resources are required beyond the counterfactual, an economic evaluation directly examines treatment contrast.

Knowing more about the counterfactual amount of resource has implications for interpreting effect sizes. A striking example here is Head Start. Evaluations of Head Start have found mixed results in terms of effectiveness. However, as described in detail in Kline and Walters (2016), the comparison group is composed one-third of children who are in public preschool programs or receive some patchwork of child care services already. Compared to these children, who are receiving resources of similar amounts to Head Start children, we might not expect Head Start to be significantly more effective. Importantly, evaluations of Head Start are not contributing to the policy debate over whether the program is worth funding; strictly, they are contributing to the policy debate over whether Head Start is preferable to another form of mixed preschool investments, where that form and mix—and its cost—is not specified in detail.

With more cost analyses being performed, the extent of differential resource use and its implications for interpreting effectiveness in terms of treatment contrast are becoming clearer. We illustrate with two additional examples.

Reading Partners is a volunteer tutor supplemental reading program for students 1.5 to 2.5 years behind grade level in reading. Experimental evidence finds that *Reading Partners* increased reading skills in Grades 2 and 5 and cost $3,610 per student (Jacob, Armstrong, Bowden, & Pan, 2016). By law, all students who are behind in reading must receive supplemental support; the counterfactual group therefore received alternative (and varied) supplemental reading programs; these cost between $1,050 to $4,890 per student. Net, the treatment contrast for *Reading Partners* is estimated at 48 extra minutes per week, and the treatment contrast in terms of resources is less than $1,000. In light

of this treatment contrast, policy makers might view reading gains much more favorably.

Another example is *Success for All*, an extensive reading intervention running through Grades K to 6. In experimental evaluation, the intervention was found to improve phonics but not reading fluency or comprehension (Quint et al., 2015, Table 5.2). Expressed in this way—resource intensive but only modestly effective—*Success for All* would not seem to be very promising. However, many schools invest in programs to improve reading. When compared against alternative reading programs at counterfactual schools, Quint et al. (2015, p. 99) estimate the resource cost at $227 higher per student per year for *Success for All*. Although not a trivial amount when aggregated across a district, it is considerably below what one might anticipate from a simple description of *Success for All*.

Evaluations of many educational interventions would benefit from a clearer statement of treatment contrast. Some interventions may be fully incremental; that is, the counterfactual group receives nothing. However, many interventions involve redistributing resources or replacing one program with another or comparing students doing one activity to those doing another. The fundamental principle of opportunity cost—valuing resources against their next best alternative use—can help education research focus on treatment contrast.

Fidelity of Implementation

Establishing implementation fidelity is a critical component of program evaluation. As described by Weiss et al. (2014, p. 783), "Implementation process influences the services that are offered and how they are delivered, which in turn influences the treatment that is received by program clients." (In our discussion of fidelity, we include any deviation in treatment across sites or subjects.) Methodological literature on program fidelity has focused on the need for valid fidelity indices and instruments (see Carroll et al., 2007; Lewis et al., 2015; Mowbray, Holter, Teague, & Bybee, 2003). However, these fidelity indices and instruments are often specific to each program or rely on researcher judgment as to what constitutes a faithful implementation. Instead, economic evaluation offers an alternative, formal way to think about program fidelity.

From an economic perspective, program fidelity across intervention sites can be defined in two ways. One definition is very simple: Program fidelity occurs when each site spends the same amount of resource per student (adjusting for local price levels). Under this definition, sites are free to use resources as appropriate, but the amount of total resource should faithfully correspond to the resources implied by the implementation design. The second definition is stricter: Program fidelity occurs when each site uses the same inputs/ingredients and remunerates them at the same rate at each site. So a mentoring intervention would have site-level fidelity if each site hired mentees with the same level of experience and qualifications and if it paid these mentees at the same rate. Under these definitions, if total spending or the pattern of spending differ significantly across sites, the program is not being implemented with fidelity.

In fact, this economic perspective corresponds to other descriptions of fidelity. Hulleman and Cordray (2009) identify dose, exposure, and quality as categories for describing fidelity. Treatments with higher doses, greater exposure, and of higher quality are likely to require more resources; and the amount of additional resource can serve as a quantifiable measure of increases in dose, exposure, or quality. Resource-based measures of program fidelity therefore complement these other descriptions. Resource-based measures of fidelity are standardized (either in dollars or as percentage deviation from resources required for fidelity); but this is useful in situations where other measures of fidelity are *ad hoc*. To illustrate the economic approach to fidelity, we draw on two examples: *Read180* and *Talent Search*.

Read180 is a literacy program for struggling readers in Grades 4 through 12. Levin et al. (2017) estimated the cost to implement *Read180* as recommended by the developers and as implemented in schools. The developers' recommended version of *Read180* was estimated to cost $1,420 per student. As actually implemented, the cost ranged from $370 up to $1,950 (25%–140% of what was recommended). Given this resource difference, it seems unlikely that *Read180* was implemented with fidelity in these schools. (If fixed costs are very large, it is mathematically possible that variable costs are equivalent even for this example of *Read180*; we appreciate a reviewer pointing this out.)

Our second example is *Talent Search*, a program to increase high school completion and college enrollment (Bowden & Belfield, 2015). *Talent Search* sites have discretion over staffing, services provided, service location, and grade levels served. In analysis across nine sites, Bowden and Belfield (2015) identify significant variations in resource use. Many sites obtain leveraged resources or use subsidized college facilities. More importantly, some sites funded students to participate for one year; others funded students who could participate over several grades. Some sites spent double the amount of federal funding; accounting for years of participation, spending per treated student ranged up to 10 times as much.

Overall, studies of implementation focus either on processes or on the technology of the intervention. As an additional way to capture fidelity, an economic approach focuses on the resources actually used. By necessity, the former approach relies on instruments and indices specific to each intervention; it is thus difficult to generalize as to what is high- or low-fidelity implementation (Mowbray et al., 2003). By contrast, by measuring in dollars the resources used, the economic approach is numerical and prescriptive, allowing implementation fidelity to be described more transparently and more formally.

The Importance of Mediators

Program effects can be understood through mediators, the links in the causal chain between treatment and targeted outcomes (Weiss et al., 2014, Figures 1 and 2). In educational research, mediators are especially important and might be the most salient part of an intervention. For example, educational vouchers are presumed to be effective because they influence which school a student attends: Educational outcomes will improve only if vouchers allow students to attend higher quality schools. Choosing a high-quality school mediates the outcome. We

argue that many educational interventions cannot be evaluated without information on the induced change in resources, that is, the higher quality school attended. (Potentially, an intervention might have many mediators even if it has a single outcome.)

Many interventions in education induce behavior change and future service uptake (Bowden, Shand, Belfield, Wang, & Levin, 2016, refer to these as "mediated service interventions"). The complete impact of the intervention depends upon placing students on a different educational or developmental pathway so they receive a different set of services. Preschool may increase earnings in adulthood (Chetty et al., 2011). But if this effect is mediated through increased resources—if the preschool child is then tracked into a higher quality school or if the parents devote more time to their child—then the earnings outcome reflects both the treatment (preschool) and the mediated services (changes in K–12 resources and parental investments). The impact of a new placement test depends not only on the test but the type of remedial services received based on the test result. The impact of a grade retention policy depends on both whether the student is retained in grade and what services are received by retained versus nonretained students. The impact of a financial aid application "nudge" is both an increase in college enrollment and an increase in resources/aid when in college (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012).

With a significant change in mediated services, formal cost analysis becomes more necessary. For a full evaluation, the cost of the mediated services should be added to the cost of the treatment. Mediated services may cost more than the treatment itself: Such a program is unlikely to be affordable. But mediated services may be negative cost, offsetting the treatment cost; such a program would be very attractive. For example, interventions to improve assignment to remediation have reduced the number of students in remediation (Scott-Clayton, Crosta, & Belfield, 2014); interventions to reduce grade retention may reduce the time students spend in school overall. In these cases, if students are no worse off on average, then the resource savings (from fewer students in remediation or retained in grade) may justify the intervention.

To illustrate service mediation interventions, we use the example of *City Connects*. Increasingly, schools are drawing on a broad array of externally-provided support services to enhance students' development (Berliner, 2009). *City Connects* places professional coordinators in schools to partner with community agencies and service providers so as to streamline student support referral and management. *City Connects* has positive effects on academic achievement of 0.38 *SD* in eighth-grade ELA and math; and the program coordinators cost $1,540 per student (Bowden et al., 2016; Walsh et al., 2014). At issue is what amount of services mediated this achievement gain. Based on interviews and documentation, Bowden et al. (2016) identified significantly more mediated support services for *City Connects* students; conservatively, the mediated services (at $3,030 per student) required twice as much resource as the program coordinator.

Overall, for some education interventions we might almost conclude that "the mediators are the message." Resources that come via mediation are a critical part of the intervention, and the success of the intervention may depend upon these as much as the initial treatment.

Power of the Test

In randomized trials in education, researchers are expected to calculate the minimum detectable effect size (MDES), that is, the smallest true effect that has a "good chance" of being found to be statistically significant (Bloom, 1995). Helpful tools are now available to ensure that power calculations for MDES are done correctly, accounting for individual, hierarchical, and block random assignment as well as for nonexperimental methods (Dong & Maynard, 2013). These power calculations specify what size effect has a good chance of being detected given the research design of the impact evaluation. The purpose is to ensure that the sample size and research design are such that valid impacts can be identified (Schochet, 2008).

In an economic evaluation, the impacts must be translated into resource use. For example, an impact evaluation might test to see if a program such as *Check and Connect* increases high school attainment; an economic evaluation would rely on the same impacts but translate them into dollars (e.g., the extra earnings from higher attainment). Researchers will perform power calculations so as to allow for identification of a valid impact on attainment. But for an economic evaluation, the requirement is a power calculation so as to allow for identification of a valid impact in dollars. These power calculations need not be the same (see Schochet, 2008). Simply, if we change what outcomes are being tested, we need to make new power calculations.

Instead of a power calculation to identify an MDES, the power calculation may need to identify a minimum detectable economic consequence (we might refer to this as "MDEC"). This MDEC is the smallest true dollar consequence that has a "good chance" of being found to be statistically significant. Of interest for cost analyses is not whether impacts *per se* are statistically significant but whether there are significant economic resource consequences derived from those impacts (see Zerbe, Davis, Garland, & Scott, 2014, pp. 370–371). It cannot be assumed *ex ante* in any given study that the same power calculations will apply to educational impacts as they do to dollar consequences. The precision standard for MDES depends on the distribution of effects; the precision standard for MDEC depends on the distribution of effects and their dollar consequences. As a general rule, we cannot say whether the MDES or the MDEC will necessitate a larger sample size (although our interpretation of Schochet [2008] is that larger samples are needed for MDEC; sample sizes to test for cost-effectiveness in health economics tend to be larger than those to test for effectiveness [Willan, 2011]).

In principle, economic considerations can be linked to the MDES. One way is to directly insert a parameter into the formula for MDES calculation that captures the economic value of the impact. The analyst would first predict the cost of a proposed intervention, then determine the size of benefit required to exceed this cost, and then determine the sample size needed to obtain a statistically significant estimate of benefit minus costs. The risk is that an impact evaluation is correctly powered for MDES but underpowered for an economic evaluation.

We illustrate the basic principle of MDEC using a study of popular literacy interventions to improve phonics. Impact studies of four leading interventions yield effect size gains per student of 0.22 to 0.34 for *Fast ForWord, Sound Partners, Wilson Reading Systems*, and *Corrective Reading* (Hollands et al., 2015). A very large sample would be needed to distinguish between these interventions (to identify an effect size gain of 0.12, for example, would require a sample of 547 for each comparison). However, these interventions cost very different amounts. Thus, they yield very different cost-effectiveness ratios: Relative to *Fast ForWord*, the cost per effect size gain was 5 times as high for *Sound Partners*, 20 times as high for *Wilson Reading Systems*, and over 60 times as high for *Corrective Reading*. To identify an economic difference between *Fast ForWord* and *Corrective Reading* would therefore involve a very different power calculation than to identify an achievement difference.

We can numerically illustrate the MDEC with a stylized, artificial example. Imagine an educational intervention is randomly assigned to 100 students with a control group of 100 students. The intervention is intended to boost attainment but it might also be evaluated in terms of adult earnings. For information on attainment and earnings, we use the 2017 Current Population Survey (CPS), artificially designating Iowan residents as the treatment group and Kansas residents as the control group. We draw 100 persons from the CPS to calculate the MDES and the MDEC (power = 0.8, alpha = 0.05, total sample size = 200).

Average (standard deviation) attainment in Kansas is 13.62 (2.1) years; the standard deviation of attainment in Iowa is 2.4. Given these parameters, the smallest mean detectable effect for an intervention in Iowa would be 0.903 years or an MDES of 0.375. Similarly, average (standard deviation) earnings for the sample of 100 Kansans is $27,960 ($20,430), and the standard deviation of earnings in Iowa is $32,880. Given these parameters, the smallest mean detectable effect for an intervention in Iowa would be $10,910 or an MDES of 0.334. Because this is in dollars, we refer to it as the MDEC.

Thus, the MDES varies with the outcome. Critically, attainment and earnings are correlated but they do not have the same distributions; and only the latter (denominated in dollars) can be compared to the costs of the intervention. If we wish to compare the costs of the intervention to its outcomes, the MDES for attainment is unlikely to be valid. In this example, the MDES and MDEC are not close: A mean detectable effect of 0.903 years is unlikely to increase anyone's earnings by $10,910 (approximately, one year of attainment adds 10% to earnings, i.e., only $3,288). The MDEC is needed to ensure a valid research design.

Power calculations are an essential step in ensuring that a research study can validly identify statistically significant differences in impact. An economic approach would extend these power calculations to ensure that a research study can validly identify significant differences in dollars or resource use. As yet, these power calculations are not codified (for impacts, see Dong & Maynard, 2013), and so we cannot predict what statistical techniques may be necessary. However, there may be nontrivial implications for research design.

Meta-Analysis

Increasingly, meta-analysis is applied to estimate the "average" effect size of an educational intervention (Borenstein et al., 2009). Meta-analyses are common across educational research (e.g., on within-class ability grouping or class size reduction, see Ahn, Ames, & Myers, 2012). They are valuable for identifying how variations in study design and population heterogeneity affect estimates of outcomes (Scammaca, Roberts, & Stuebing, 2014). However, when we apply cost-based analysis, there are two important concerns in relation to meta-analysis.

The first concern is that a meta-analytic estimate cannot simply be linked to a measure of costs or resource use. Therefore, it may not be possible to use meta-analytic estimates in a CEA. CEA requires linkage of the inputs used for specific programs to their outcomes. Meta-analysis provides estimates from different versions of a single class of interventions to adjudicate on whether a general class of programs is effective on average. Policy makers might look at a meta-analytic effect size and ask, Is this effect size worth spending $X on? That question might not be answerable. The effect size represents the average from a "general class of programs." The cost of this general class might be unknown (for example, what does within-class ability grouping cost?). If the studies included in the meta-analysis all refer to precisely the same intervention, it may be possible to estimate the average cost; this seems unlikely.

A second concern relates to the meta-analytic method. Meta-analytic estimates are adjusted for attributes of the research evaluation. These moderator variables are typically grouped into categories defined as units, treatment, observing operations, setting, and method (Ahn et al., 2012; Cooper, 2009). However, some of these characteristics are almost certainly correlated with the costs of an intervention. The "units" category may refer to the grades of the students and the scale of the intervention: Typically, more resources are allocated to students in higher grades and to larger scale programs. The "treatment" category may refer to the duration of the intervention; typically, longer programs have more resources allocated to them. The "setting" may refer to the locality of the intervention; typically, interventions that are off-site from a school require more resources. Controlling for these categories is therefore, to some extent, adjusting for resource use; and the extent of this adjustment is unknown. (One solution is to report meta-analytic estimates that vary across cost factors; we appreciate this suggestion from a reviewer.) To illustrate this argument, we examine meta-analytic research on the association between computer-aided instruction and learning outcomes.

Several meta-analytic studies and systematic reviews identify a learning advantage for students who receive computer-aided instruction (U.S. Department of Education, 2009). Sosa, Berger, Saw, and Mary (2011) estimates an average effect size gain of 0.33 standard deviations after extensive moderator analysis. However, many of these moderators are related to the costs of the intervention in each case. For example, one moderator is if the computer-aided instruction included additional time; unsurprisingly, the effect size gain is five times larger in studies where additional time is allowed. Similarly, when the computer-aided

tool is a supplement the effect size gain is more than twice as large. Other economically relevant moderators include if the instruction was assessed and if the instruction included features such as web communication or number cruncher tools. Each of these moderators is related to the amount of resource used for computer-aided instruction.

For meta-analysis, an economic approach can provide guidance on which moderators to consider and how moderators may be interpreted. As a simple assumption, moderators might be understood in terms of the additional resources they represent; more resource-intensive moderators should be associated with greater average effectiveness.

Summary

Our focus in this essay has been on how economic methods may complement and so enhance impact evaluations. Of course, economic evaluations face many methodological challenges of their own, and economic analysis is not necessarily determinative (Farrow & Zerbe, 2013). However, impact evaluations and economic evaluations are linked, and this linkage is often neglected. One consequence is that economic evaluations become methodologically more challenging (because some outcomes have no clear economic interpretation or because effect sizes have already adjusted for resource use). As economists, our worry here is that an economic evaluation will not be possible. But another consequence, the one we emphasize here, is that impact evaluations are themselves less valid or less useful. As education researchers, our worry here is that our analysis and evidence will not be sufficiently helpful for decision makers.

REFERENCES

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, *82*, 436–476.

Bailey, D., Duncan, G. J., & Watts, T. W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39.

Belfield, C. R., Bowden, A. B., Klapp, A., Levin, H. M., Shand, R., & Zander, S. (2015). The economic value of social and emotional learning. *Journal of Benefit-Cost Analysis*, *6*(3), 508–544.

Berliner, D. C. (2009). *Poverty and potential: Out-of-school factors in school success*. Boulder, CO; and Tempe, AZ: Education and the Public Interest Center & Education Policy Research Unit. Retrieved from http://epicpolicy.org/publication/poverty-and-potential

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment. *Quarterly Journal of Economics*, *127*(3), 1205–1242.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power. *Evaluation Review*, *19*(5), 547–556.

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2018). *Cost-benefit analysis: Concepts and practice* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: Wiley & Sons.

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

Bowden, A. B., Shand, R., Belfield, C. R., Wang, A., & Levin, H. M. (2016). Evaluating educational interventions that induce service receipt: A case study application of City Connects. *American Journal of Evaluation*, *23*(1), 1–16.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*, 40. doi:10.1186/1748-5908-2-40.

Castleman, B. L., Page, L. C., & Schooley, K. (2014). The forgotten summer: Does the offer of college counseling after high school mitigate summer melt among college-intending, low-income high school graduates? *Journal of Policy Analysis and Management*, *33*(2), 320–344.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Whitmore Schanzenbach, D., & Yagan, D. (2011). How does your kindergarten class affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*(4), 1593–1660.

Cooper, H. (2009). *Research synthesis and meta-analysis: A step-by-step approach*. Thousand Oaks, CA: SAGE Publications.

Dong, N., & Maynard, R. (2013). PowerUp! A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67.

Duckworth, A. L., & Yeager, S. (2015, May). Measurement matters: Assessing qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*, 405–432.

Farrow, R. S., & Zerbe, R. O. (Eds.). (2013). *Principles and standards for benefit-cost analysis*. Cheltenham, UK: Edward Elgar.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*, 451–464.

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness*, *9*(1), 3–53.

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.

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

Kline, P., & Walters, C. (2016). Evaluating public programs with close substitutes: The case of Head Start. *Quarterly Journal of Economics*, *131*(4), 1795–1848.

Levin, H. M., McEwan, P. J., Belfield, C. R., Bowden, A. B., & Shand, R. D. (2017). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science*, *10*, 155. doi:10.1186/s13012-015-0342-x.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340.

Nores, M., Belfield, C. R., Barnett, W. S., & Schweinhart, L. (2006). Updating the economic impacts of the High/Scope Perry Preschool program. *Educational Evaluation and Policy Analysis*, *27*, 245–261.

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence*, *53*(Nov-Dec), 202–210.

Quint, J., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *Scaling up the success for all model of school reform*. New York, NY: MDRC.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*(3), 328–364.

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

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*(3), 371–393.

Sklad, M., Diekstra, R., De Ritter, M., Jehonathan, B., & Gravestein, C. (2012). Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools*, *49*, 892–910.

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.

Sosa, G. W., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research*, *81*(1), 97–128.

U.S. Department of Education. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Washington, DC: Office of Planning, Evaluation, and Policy Development.

Walsh, M. E., Madaus, G. F., Raczek, A. E., Dearing, E., Foley, C., An, C., . . .Beaton, A. (2014). A new model for student support in high-poverty urban elementary schools: Effects on elementary and middle school academic outcomes. *American Educational Research Journal*, *51*(4), 704–737.

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*(4), 778–808.

Willan, A. R. (2011). Sample size determination for cost-effectiveness trials. *Pharmacoeconomics*, *29*, 933–949.

Zerbe, R. O., Davis, T. B., Garland, N., & Scott, T. (2013). Conclusion: Principles and standards for benefit-cost analysis. In S. Farrow & R. O. Zerbe (Eds.), *Principles and standards for benefit-cost analysis* (pp. 364–445). Northampton, MA: Edward Elgar.

AUTHORS

CLIVE R. BELFIELD is a professor of economics at Queens College, City University of New York, 65-30 Kissena Boulevard, Flushing, NY 11367; *belfield@qc.edu*. His research focuses on economic evaluation of education interventions and programs.

A. BROOKS BOWDEN is an assistant professor of methods and policy at North Carolina State University, 2310 Stinson Hall, Raleigh, NC 27695; *brooks_bowden@ncsu.edu*. Her research focuses on the methods of economic evaluation and applications of those methods to programs that address poverty in the classroom.

Manuscript received January 25, 2018
Revisions received July 2, 2018, and October 13, 2018
Accepted October 22, 2018