JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2016, VOL. 9, NO. 1, 30–53
http://dx.doi.org/10.1080/19345747.2015.1055639

Cost-Effectiveness Analysis of Early Reading Programs: A Demonstration With Recommendations for Future Research

Fiona M. Hollands[a], Michael J. Kieffer[b], Robert Shand[a], Yilin Pan[a], Henan Cheng[a], and Henry M. Levin[a]

ABSTRACT

We review the value of cost-effectiveness analysis for evaluation and decision making with respect to educational programs and discuss its application to early reading interventions. We describe the conditions for a rigorous cost-effectiveness analysis and illustrate the challenges of applying the method in practice, providing examples of programs for which we have estimated costs, but find effectiveness data lacking in comparability. We provide a demonstration of how cost-effectiveness analysis can be applied to two early reading programs: the Wilson Reading System and Corrective Reading. We use existing effectiveness data from an experimental evaluation in which the programs were implemented under similar conditions and the use of common outcomes measures for both programs yielded data that are comparable. We combine these data with cost data we collected using the ingredients method to calculate cost-effectiveness ratios for the alphabetics domain. A complete picture of the relative cost-effectiveness of each program could be provided if effectiveness metrics were available for fluency, vocabulary, and comprehension. We highlight the obstacles to applying cost-effectiveness analysis more frequently and recommend strategies for improving the availability of the requisite data.

KEYWORDS

cost-effectiveness
program evaluation
early reading

Almost one trillion dollars or 15% of total government expenditure is devoted to education in the United States (U.S. Government, 2013). Despite this enormous investment, policymakers rarely choose educational programs based on an assessment of costs and cost-effectiveness. When they do make decisions based on evidence, they typically rely only on evidence of effectiveness. Tsang (1997), Levin (2001, 2011), Levin and Belfield (2015), Levin and McEwan (2001), and Harris (2009), have argued that both costs and effects must be evaluated when considering educational interventions. By selecting programs with the highest effectiveness relative to their costs, education decision makers could improve the productivity of education, perhaps by large magnitudes.

Rigorous experimental and quasi-experimental studies are increasingly sponsored to assess the effectiveness of educational programs, but these rarely address the resources required for effective implementation and their associated costs. Although a method for

CONTACT Fiona M. Hollands fmh7@tc.columbia.edu Teachers College, Columbia University, Education Policy and Social Analysis, 525 West 120th Street, Box 181, New York, NY 10027, USA.
[a]Teachers College, Columbia University, New York, New York, USA
[b]New York University, New York, New York, USA

conducting cost-effectiveness analysis in education has been available for many years using straightforward cost accounting methods based upon opportunity costs (see Chambers & Parrish, 1994a, 1994b; Levin, 1975), its application has been limited to only a few educational interventions such as teacher selection (e.g., Levin, 1970); computer-assisted instruction, smaller class sizes, longer school days, and cross-age tutoring (e.g., Levin, Glass, & Meister, 1987); adolescent literacy programs (e.g., Kim et al. 2011; Levin, Catlin, & Elson, 2007); and dropout prevention (e.g., Hollands et al., 2014).

Despite the fact that over one third of the elementary school day is devoted to English, reading, and language arts (U.S. Department of Education, 1997), a percentage of time that has remained relatively stable on subsequent administrations of the Schools and Staffing Survey through 2007–2008,[1] cost-effectiveness analysis has rarely been applied to early reading programs. With average total expenditures per student in U.S. public schools at $11,153 in 2010–2011 (National Center for Education Statistics, 2014), spending on literacy is approximately $3,800 per student per year. Results for U.S. students' literacy performance are, nevertheless, mediocre, with 32% of fourth-grade students scoring below a basic level of proficiency in reading as measured by National Assessment of Educational Proficiency tests.[2] Poor reading skills often lead to grade retention (Bowman-Perrott, Herrera, & Murry, 2010; Jimerson & Kaufman, 2003), which adds the costs of a full year of education. Supplemental remedial reading programs for struggling readers are also costly (e.g., Every Child a Chance Trust, 2009; Hollands et al., 2013; Rouse & Krueger, 2004; Simon, 2011). When reading issues are severe enough to merit special education, district costs per student may double (Chambers, Parrish, & Harr, 2004).

Despite the strong research base available to guide effective early reading instruction, efforts to implement effective approaches on a large scale have been disappointing (e.g., Duke & Block, 2012; Gamse, Jacob, Horst, Bouley, & Unlu, 2008), suggesting a gap between evidence and implementation. Potential explanations for this gap include the uncertainty of the cost requirements for implementing early reading programs, and the suboptimal allocation of resources. If the full costs of implementation were considered in the initial decision-making process for the selection of reading programs, we expect that fewer problems would be encountered later as a result of insufficient resources being dedicated to ensure that the chosen program is implemented with fidelity. It is possible that inefficient decisions are being made by investing limited resources in programs that are not the most cost-effective.

The implementation requirements for effective programs to teach early reading can vary greatly in their implications for costs. Program duration varies from several weeks to several years. Some programs target entire classrooms of students, while others target small groups or individual students. Certain programs require highly trained, full-time teachers, while others use part-time paraprofessionals. Some programs require the purchase of expensive computer software or other materials, while others require only inexpensive teachers' manuals. For reading specialists, curriculum leaders, and other education decision makers to make informed decisions when selecting early reading programs, information about the potential alternatives must include attention to costs as well as effects. Programs that are excessively costly as designed may be less likely to be implemented fully or faithfully. In experimental trials, the costs of implementation are typically borne by the researchers,

[1] http://nces.ed.gov/surveys/sass/tables/sass0708_005_t1n.asp
[2] http://nationsreportcard.gov/reading_math_2013/#/what-knowledge

obscuring the extent to which real-world implementation may deviate from developers' recommendations due to concerns about costs. The result is that effects obtained in the field may be smaller than those suggested by experimental trials.

Combining cost and effectiveness data on alternative interventions can indicate which interventions are likely to provide the best results for a given level of investment. However, few empirical studies have been conducted on the cost-effectiveness of reading programs (Hummel-Rossi & Ashdown, 2010; Levin, 2011). Borman and Hewes (2002) compared the effects and costs of Success for All with three other interventions (Perry Preschool, the Abecedarian Project, and the Tennessee Class-Size Experiment), but all these interventions are whole-school reforms rather than reading programs *per se*. Massoni and Vergnaud (2012) present a cost-effectiveness analysis of an early literacy program, Action Lecture, used in French nursery and primary schools. However, only the teacher's salary is considered as a cost and the comparison program is class-size reduction as opposed to alternative literacy programs.

A number of other studies mention costs or cost-effectiveness of specific reading programs but do not systematically estimate both costs and effects and compare them across reading programs. For example, Rouse and Krueger (2004) provide a rough estimate of the costs of Fast ForWord, arriving at a sum of $770 per student excluding the costs of space, and suggest that the program may be cost-effective for districts in which the program can appropriately be used for many students. However, the cost estimate is not based on documentation of actual resource use in a particular implementation, no cost-effectiveness ratio is provided, and the program is not compared with any alternative intervention or with "business-as-usual." Every Child a Chance Trust (2009) presents a cost-benefit analysis of Reading Recovery, estimating program costs at around $4,600 per student in 2008 dollars and weighing these against the long-term costs of failing to prevent reading difficulties, but no comparison to alternatives is provided.

Other researchers have considered reasons why one particular early literacy intervention may be more cost-effective than another, but they have not focused on collecting detailed cost estimates to support these claims and do not estimate cost-effectiveness ratios to compare the programs. For example, Hatcher et al. (2006) compare two programs derived from Reading Recovery and used in the UK. They suggest that the Early Literacy Support Programme is more cost-effective than the Reading Intervention Programme on the basis that it can be delivered using fewer hours of a teaching assistant's time. Justice, Kaderavek, Fan, Sofka, & Hunt (2009) suggest that a print referencing intervention is cost-effective compared to other early literacy interventions due to the low costs of the storybooks required to implement this approach and the relatively limited training or support needed for teachers to deliver it. Pikulski (1994) reviewed five reading programs used with first-grade children, including Reading Recovery, and postulates that programs that prevent reading difficulties are "very cost effective when compared against the costs involved in remedial efforts" (p.30).

One notable exception in the area of early reading in which the author purposely set out to compare cost-effectiveness of several programs is Simon's (2011) analysis of Classwide Peer Tutoring, Reading Recovery, Success for All, and Accelerated Reader. Combining effectiveness data from existing studies with cost data collected retrospectively, Simon found significant differences across the four programs in costs ($500–$11,700 per student per year), and cost-effectiveness ($1,400–$45,000 per unit increase in effect size for reading outcomes).

Even this limited evidence base on costs and cost-effectiveness of reading programs suggests that decision makers could deploy resources more efficiently when selecting programs to improve reading.

In what follows, we describe our efforts to obtain existing effectiveness data on early reading programs for the purposes of combining them with our own cost estimates of the programs to conduct cost-effectiveness analyses. We estimated costs of reading interventions using the "ingredients method" of cost accounting, applying a uniform cost method across all programs. We obtained effectiveness data from the summary of effectiveness of educational interventions provided by the What Works Clearinghouse (WWC). The WWC was initiated by the U.S. Department of Education to assess the evidence supporting effectiveness of educational interventions by reviewing the validity of evaluations conducted by researchers. Unexpected limitations in the availability of comparable effectiveness data on multiple reading outcomes for each reading program listed as "effective" restricted us to a comparison of just two reading programs for which the effectiveness data were adequately comparable. We use these programs to demonstrate the method of cost-effectiveness analysis and explain the obstacles we encountered in attempting additional analyses with other reading programs for which we were successfully able to estimate costs. We recommend strategies for improving the availability of the requisite effectiveness data so that cost-effectiveness analysis can be conducted more often to facilitate efficient decision making.

Effectiveness of Early Reading Programs

Reading researchers have made considerable progress in the last thirty years in identifying the elements of effective early reading instruction. Synthesizing evidence from a large body of research, the National Reading Panel report (National Institute of Child Health and Human Development, 2000), the National Research Council report on preventing reading difficulties (Snow, Burns, & Griffin, 1998), and other seminal reviews (Adams, 1990; Ehri, Nunes, Stahl, & Willows, 2001; Ehri, Nunes, Willows, et al., 2001) provide valuable recommendations for how to teach children to read in kindergarten to Grade 3. These research-based practices have informed the development of early reading programs, including curricula for classwide instruction as well as interventions for struggling readers.

Informed by these national reports, the U.S. Department of Education's What Works Clearinghouse (WWC) identified four domains of interest for early reading: alphabetics (i.e., skills involved in the representation of spoken sounds by letter patterns, including phonological awareness, letter identification, print awareness, and phonics), text reading fluency, comprehension, and general reading achievement (WWC, 2012). The WWC classifies word reading efficiency measures under the alphabetics domain as distinct from text reading fluency; although one could make arguments either way for where these word-level speed measures could be classified, we have followed the WWC's practice in this regard. We focus primarily on the alphabetics domain, in part because it is necessary for proficient reading (NICHD, 2000) and in part because it is the only domain for which comparable data were available from WWC for multiple programs. We acknowledge that success in the alphabetics domain is far from sufficient for proficient reading and that information on text fluency and comprehension outcomes, if the requisite effectiveness data were available, would be more valuable to decision makers.

Although the four literacy outcome domains defined by WWC are not independent but rather interrelated developmentally in complex ways (e.g., Scarborough, 2001), they provide a useful means of comparing the effects of reading programs on outcomes to which education decision makers can easily relate. The WWC reviews evidence from experimental and quasi-experimental studies of early reading programs and assigns ratings of effectiveness for individual programs based on the robustness of the evaluation study designs and their findings (see WWC, 2012; WWC, 2013). Although literacy experts do not all agree with the WWC's analysis and presentation of reading program effectiveness (e.g., McArthur, 2008; Slavin, 2008; Slavin & Madden, 2011; Slavin & Smith, 2009; Stockard, 2008; Stockard & Wood, 2013), it can provide an initial basis for guiding the selection of programs with evidence of effectiveness. Many of these researchers' concerns focus on analysis of individual programs, but some of the concerns raised regarding effectiveness data can be compounded by a need for comparable data among several programs when making cost-effectiveness comparisons.

Reading researchers have often come to conclusions about the effectiveness of programs or general approaches to teaching reading based on meta-analytic estimates in key syntheses over the past 30 years (e.g., Ehri, Nunes, Stahl, et al., 2001; Elleman, Lindo, Morphy, & Compton, 2009; Lonigan & Shanahan, 2009; NICHD, 2000; Stahl & Fairbanks, 1986). Although such estimates provide valuable information about the average effectiveness of an approach across multiple implementations and studies, they preclude matching of resource requirements and costs to a specific implementation, which is necessary for an accurate cost-effectiveness analysis. Ideally, for a policymaker to make a decision about whether to use a program in his or her school context, he or she would consult the costs and effects derived from studies that include a population of students similar to the decision maker's intended audience. Decision making could be further improved if more studies compared the effects of alternative reading programs as opposed to simply assessing whether one program is more effective than business-as-usual.

Comparability of Effectiveness Data on Early Reading Interventions

The first task in setting up a cost-effectiveness analysis that can provide useful information for decision makers is to identify programs that can serve as viable alternatives in addressing the same outcomes and that have been evaluated in ways that yield comparable effectiveness estimates. Reading programs are designed to improve a variety of early reading domains and constructs within those domains so that a fair and useful comparison must start with programs addressing the same specific domains and constructs. Finding comparable programs can be difficult because evaluation studies do not necessarily measure all constructs targeted by a program, and sometimes employ measures to assess impact on constructs that a program does not aim to address. Such inconsistencies and gaps in measurement of effects are problematic when attempting to compare the overall benefits of a program. Even when programs do address the same broad reading domain, the measures used to capture program impacts are often not identical across studies. Some measures are more sensitive to instruction than others, or may be considered "treatment-inherent" as opposed to "treatment-independent" (Slavin & Madden, 2009). Ideally, a cost-effectiveness analysis would compare

programs for which effectiveness has been assessed for multiple reading outcomes using the same measures.

A second consideration for comparability relates to grade spans and student abilities because the target population for early reading programs varies and generalizability of the results to other populations may be inappropriate. It is well documented that the rate of gain in reading skills slows as students ascend grade levels (e.g., Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Kieffer, 2011; Roberts, Mohammed, & Vaughn, 2010). Hill, Bloom, Black, and Lipsey (2007) report that annual normative gains in reading test scores vary substantially but predictably across grades: when expressed as effect sizes, the average gain in reading tests scores for students progressing through kindergarten to first grade is 1.52, compared with 0.97 for Grade 1–2 students, and 0.36 for Grade 3–4 students. Programs targeting older students may therefore appear less effective than those targeting younger students, even if the students show a greater than typical gain in reading. This suggests that a useful comparison of reading programs would only include studies of programs involving students of the same age.[3]

The Challenge of Matching Cost Data With Effectiveness Data

Cost data are rarely collected simultaneously with effectiveness data and few, if any, study reports provide substantial details regarding implementation. As a result, it is usually necessary to collect data on resource requirements retrospectively. This requires reconstructing program implementation through historical documents and interviewing personnel involved in actual program implementation. The latter is only feasible when knowledgeable personnel can be identified and can accurately recall the necessary details. In conducting cost analyses, we find that accuracy of recollection drops markedly beyond about five years and can be quite poor beyond ten years. To assure the greatest possible accuracy in retrospective collection of resource utilization, costs of program implementation must be based on fairly recent evaluation studies. In the specific case of early reading instruction, many high-quality evaluations were conducted more than a decade ago (e.g., Foorman, Francis, Fletcher, Schnatschneider, & Mehta, 1998), preventing their use in this type of analysis.

To further assure accuracy of cost estimates, costs of a program must be determined for a specific implementation for which details of resource use are obtainable. Just as different program sites often show different levels of effectiveness in impact studies, Levin, Catlin, and Elson (2007), and Levin et al. (2012) have clearly demonstrated that different implementations of the same program utilize varying amounts of resources, which in turn result in varying costs. Specifically, Levin et al. (2007) found that a literacy program, READ 180, that is intended to be implemented uniformly across sites actually varies in implementation costs from $285 to $1,514 per student. Levin et al. (2012) found site-level costs for Fast ForWord Reading 1 at one site to be 36% lower than at another. This supports a strategy of matching costs of a particular implementation with the effects of that specific implementation. However, effectiveness of any single early reading program has in many cases been assessed through multiple studies over the years and researchers frequently use meta-analytic

[3] It should be noted that effect sizes are more useful for comparison purposes than for direct interpretation, because they are relative measures without units. Under a normal distribution, an effect size of 1 represents a substantial increase: movement from the 50th percentile on the underlying measurement instrument to about the 84th percentile.

methods to combine results of the various studies to provide an effect size that reflects the average effect of an instructional approach or type of instruction, averaging out variations across the studies (e.g., Elleman et al., 2009; NICHD, 2000; Stahl & Fairbanks, 1986). Although this approach may present a more stable picture of program performance across multiple implementations of the program than any single instance, the value of a cost-effectiveness analysis is that it allows decision makers to identify exactly what resources were used to obtain a particular level of impact. For example, implementing a program with trained reading specialists may result in greater impact but also be more expensive than the same program implemented with parent volunteers. The decision maker needs to know whether he or she can afford the trained specialists or even has access to the necessary personnel. Hence, while an average effect size for a program may be helpful to choose among programs from the perspective of expected impact, it does not reflect the resources required to obtain that impact in practice.

To promote attention to costs and cost-effectiveness analysis of early reading programs, we demonstrate here how a uniform cost accounting approach can be combined with existing effectiveness data to conduct retrospective cost-effectiveness analyses of two early reading programs: the Wilson Reading System and Corrective Reading. These programs were evaluated for effectiveness by Torgesen et al. (2006) in the same randomized controlled trial, with third-grade students of similar reading abilities, using the same "dosage," and measuring impact on reading outcomes using the same instruments. This situation provides effectiveness results that are comparable in all respects. However, we note that such experiments are rare and are often conducted under "superrealized" (Cronbach et al., 1980) circumstances that do not reflect how programs are typically implemented in schools. We also describe challenges we encountered in finding suitably comparable effectiveness data that would allow for a cost-effectiveness analysis of a larger number of programs, which was our original goal. We examine how education decision makers can use the results of cost-effectiveness analysis and how future research can be designed to more routinely capture considerations of cost-effectiveness.

Methods

Selection of Reading Programs to Compare

A cost-effectiveness comparison requires two or more alternative interventions with similar goals and outcomes measured by similar criteria. Because evaluations listed in the WWC are each undertaken independently by external evaluators, they vary in terms of the specific aspects of reading assessed as well as in the choice of evaluation instruments and metrics, therefore limiting direct comparisons between the programs. We searched among the 32 programs listed in the WWC's Beginning Reading (Pre-K–3) subcategory of interventions to identify programs that met two criteria. First, they needed to show positive or potentially positive effects on reading measures that were comparable in terms of the reading outcomes addressed. Second, they needed to have at least one evaluation study published since 2005 so that the implementers could remember details of personnel and resources utilized. As noted earlier, this requirement can preclude the inclusion of important earlier studies. It may also result in a focus on recent studies that may not be representative of findings across time. Among the

programs that met our initial criteria, we aimed to select two or more that could be compared in a cost-effectiveness analysis that met the methodological requirements.

One study meeting the WWC evidence criteria, Torgesen et al. (2006), simultaneously evaluated four reading programs serving students in the third grade in a randomized controlled trial conducted during the 2003–2004 school year. This presented a potentially ideal situation for applying cost-effectiveness analysis because the programs targeted students of the same grade level and reading ability (students scoring mostly in the bottom 25th percentile on standardized tests), were implemented over the same period of time using the same dosage, and impact was assessed using the same outcome measures. Two of the programs evaluated, the Wilson Reading System and Corrective Reading, were modified to specifically target word-level skills by eliminating the vocabulary and reading comprehension components. The effects of both programs on alphabetics were assessed using the same four measures: WRMT-R Word Identification subtest (a measure of word reading accuracy), WRMT-R Word Attack subtest (a measure of decoding accuracy), TOWRE Sight Word Efficiency subtest (a measure of word reading efficiency), and TOWRE Phonemic Decoding subtest (a measure of decoding efficiency).

The other two programs evaluated, Spell Read Phonological Auditory Training and Failure Free Reading, were intended to target reading comprehension and vocabulary. Although the reading comprehension and vocabulary outcomes are arguably more comprehensive, long-term reading outcomes than word-level skills, neither program showed impacts on the intended outcomes so we focused on Corrective Reading and the Wilson Reading System, each of which showed positive impacts on some word-level or alphabetics skills. However, the decision by Torgesen et al. (2006) to excise the reading comprehension and vocabulary components of both programs limits our ability to assess their cost-effectiveness for all important reading outcomes. Furthermore, we note that the study was conducted with small treatment samples of around 40–50 students per program, raising some concerns about generalizability of results and the propensity for small sample studies to report larger effect sizes (see Slavin & Smith, 2009).

In practice, it is rare that early reading programs are compared side-by-side under such similar circumstances. Furthermore, the implementation of randomized controlled trials often involves monitoring practices for consistency, including more frequent testing and observations than typical in order to maintain high fidelity of implementation. In the Torgesen et al. (2006) study, the Wilson Reading System and Corrective Reading were modified to fit with the planned length of the experiment (28 weeks), while in regular school practice they are delivered over two to three years. This situation clearly demonstrates the tension between designing rigorous evaluations of programs and assessing how they actually work in practice. Decision makers who are choosing among alternative programs to implement in schools must balance methodological rigor with results that can be reasonably expected in the field.

More commonly, a single reading program is evaluated and its impact is compared with business-as-usual. Illustrating the challenges that arise in conducting comparative cost-effectiveness analysis, while we were able to estimate the costs for five other programs listed by the WWC as having positive or potentially positive alphabetics outcomes, we were not able to compare them in cost-effectiveness analyses due to various concerns with the comparability of the effectiveness data available from studies that

met our initial criteria. In the first instance, Hill et al. (2007) highlight the importance of only comparing effects for students at the same grade level so we only considered comparing programs serving students of the same grade level and of approximately the same reading ability. One potentially eligible study of Fast ForWord Reading 1 (Scientific Learning Corporation, 2005) mixed effectiveness results for students in Grades 1 and 2, such that it was hard to compare with another potentially eligible study of Reading Recovery that served only first-grade students (Schwartz, 2005). We identified three potentially eligible studies of programs serving kindergarten students, but each one served a different target audience: K-PALS targeted average readers (Stein et al., 2008); Sound Partners targeted readers in the 20th−30th percentile (Vadasy & Sanders, 2008); and Stepping Stones to Literacy targeted struggling readers with behavioral disorders (Nelson, Benner, & Gonzales, 2005; Nelson, Stage, Epstein, & Pierce, 2005).

Additional comparability issues arose due to length of implementation: some programs were implemented for only five or six weeks (Stepping Stones to Literacy, Fast ForWord Reading 1) while others lasted 18–20 weeks (K-PALS, Sound Partners, Reading Recovery). Because treatment duration is likely to influence both effect sizes and costs, this mismatch in length of intervention diminishes program comparability in a cost-effectiveness analysis. Other limitations included the fact that two potentially eligible studies (those of K-PALS and of Reading Recovery) used a measure created by the program developers to determine effects in alphabetics while others used independently developed measures. Slavin and Madden (2009) demonstrate that treatment-inherent measures produce much larger positive effect sizes than treatment-independent measures. Additionally, there is evidence that when developers evaluate their own programs, as was the case in many of the studies we identified, effect sizes reported are higher than for programs evaluated independently (Petrosino & Soydan, 2005).

To provide a demonstration of the cost-effectiveness method applied to Wilson Reading System and Corrective Reading, we describe each program in the following sections, briefly reviewing evidence of program effectiveness and details of the Torgesen et al. (2006) study that we used for both effectiveness data and to estimate the costs of implementation. We summarize the details of program implementation and the effectiveness data in Table 1.

Corrective Reading

Corrective Reading is a remedial reading program that serves students in Grade 3 or above who are reading below their grade level. It is delivered as a pull-out program to small groups of students or to whole classrooms. The program employs a direct instruction approach with a trained teacher delivering a scripted presentation at a brisk pace and engaging the students with exercises and examples. It consists of two strands, decoding and comprehension. The developer recommends 240 hours of instruction delivered in 45-minute sessions four to five times per week over a period of two to three years. WWC (2007a) concluded that Corrective Reading has potentially positive effects for alphabetics and fluency, and no significant effects for comprehension based on one study (Torgesen et al., 2006) of 79 students. Effect sizes reported for third-grade students were 0.22 for alphabetics, 0.27 for fluency, and 0.17 (*ns*) for comprehension. Although this study included both third and fifth graders, we estimated costs only for the third graders, consistent with our focus on early reading. Ten trained teachers delivered instruction to groups of three third graders, one hour per day, five days per week over 28 weeks in 14 schools in Pennsylvania. On average, treatment students

Table 1. Program details and effect sizes observed for the Wilson Reading System and Corrective Reading in Torgesen et al. (2006) study.

	Corrective Reading	Wilson Reading System
Program/study characteristic		
Grade level of students in study	3	3
Targeted students in study	Bottom 25th percentile of readers	Bottom 25th percentile of readers
Total number of students receiving intervention	44 across 11 schools	53 across 10 schools
Duration	28 weeks	28 weeks
Point of impact testing after program start	28 weeks	28 weeks
Dosage	60 mins/day, 5 days/week	60 mins/day, 5 days/week
Delivery	1–3 pull-out with Corrective Reading teacher, supplements classroom instruction	1–3 pull-out with Wilson Reading System teacher, supplements classroom instruction
Effect Sizes		
Alphabetics (average effect size)†	0.22❖	0.33❖
Text Reading Fluency††	0.27*	0.15*ns*
Comprehension (average effect size)‡	0.17*ns*	0.17*ns*

Note. *Statistically significant; ❖ = this effect size is an average of four effect sizes, two of which are statistically significant; *ns* = not significant; † Measures used: TOWRE: Phonetic Decoding Efficiency; TOWRE: Sight Word Efficiency; WRMT-R: Word Identification; WRMT-R: Word Attack subtest; †† Measure used: Edformation Oral Fluency Assessment; ‡ Measures used: GRADE: Passage Comprehension; WRMT-R: Passage Comprehension.

in the study received 90 hours of instruction, far short of the 240 hours intended by the program developer. Instructional emphasis was greater on alphabetics than fluency or comprehension.

Wilson Reading System

The Wilson Reading System is a supplemental remedial reading and writing program for students in Grade 2 and above. It uses a direct, multisensory approach based on Orton-Gillingham principles (Ritchey & Goeke, 2006). A certified instructor delivers instruction to groups of one to six students, three to five times per week for 60–90 minutes. The entire 12-step curriculum can take two to three years to complete depending on the frequency of delivery. Based on the same study by Torgesen et al. (2006), WWC (2007b) reported the following effect sizes for third graders: 0.33 for alphabetics, 0.15 (*ns*) for fluency, and 0.17 (*ns*) for comprehension. The program was delivered without the vocabulary and comprehension components to groups of three third-grade students over 28 weeks at 10 schools in Pennsylvania. Sessions occurred five times per week for no longer than 60 minutes each. Seventy-one students participated in this evaluation.

Collecting Cost Data Using the Ingredients Method

We applied the ingredients method (see Levin & McEwan, 2001) to the selected early reading programs in order to calculate costs. The purpose behind the ingredients (or resource) approach is to identify all of the resources utilized in the implementation of a program and subsequently account for their opportunity costs: the value of the resources used for a program estimated by the foregone next-best alternative use, which is generally captured in the market price. This approach begins not with a budget, but with the details of the intervention and its resource requirements. Budgets provide inaccurate estimates of costs, usually

understating them (see Levin and McEwan, 2001, pp. 45–46). For example, they do not include the costs of items used in program implementation that were purchased in years prior to program operation, or that are contributed by another agency such as the state, a private institution, parents, or volunteers. They also do not amortize the costs of capital items that can be spread over many years. Additionally, budgets often list items by function (e.g., administration, instruction, professional development, training) or by "object" (e.g., teachers, substitutes, administrators), rather than by program, so that it is difficult to determine what portion of costs is attributable to which activity. Finally, budgets generally represent plans for resource allocation rather than actual costs incurred.

The aim of our cost analyses is to estimate the cost of replicating the specific implementation of each early reading program that was used to achieve the impact results observed in the selected evaluation. In the Torgesen et al. (2006) study, most or all of the costs of implementing the program were borne by the funding agency sponsoring the study so that the program was apparently "free" to the schools. However, the true cost of each program is determined by the value of the resources that are required, not by how the program is financed. Our criterion is the cost of replicating program implementation from the perspective of the typical school; therefore, we do not include the costs that were incurred in developing these programs, because they are sunk costs that would not be borne in replication. We expect that, in typical situations, most of the costs of school-based early reading programs will be borne by the school itself, while some costs—for example, a districtwide literacy coach—might be underwritten by the school district. A small percentage of costs may accrue to families in the form of volunteer time or provision of home-based reading materials. We consider only program costs above and beyond the resources students already receive as part of their regular instruction in school, that is, we identify the incremental costs of introducing the programs into existing school activities.

The programs we studied partially replaced some regular classroom reading instruction for the students receiving the intervention. In these situations where a few students were pulled out of the primary classroom to participate in a supplementary reading program, we assumed that there were unlikely to be any significant changes in instruction in the main classroom from which they were temporarily removed and hence no significant cost savings.

An initial list of the ingredients required to implement the Wilson Reading System and Corrective Reading was compiled through careful review of the Torgesen et al. (2006) study and other publicly available articles, reports, websites, and materials for each program. Subsequently, a detailed interview protocol was developed for each program to elicit further information regarding the ingredients. Because personnel typically account for 70%–80% of the costs of educational interventions (Levin, 1975), most of our interview questions sought to elicit details about the staffing involved in implementing the program, whether directly or peripherally. For example, while the evaluation report may have indicated that teachers were employed to deliver a program five times per week in one-hour sessions, we collected information regarding the qualifications and work experience of the teachers, what proportion of their work time was spent on the program, and how many hours were spent in training, preparing lessons, tracking student progress, and communicating with the classroom teacher, principal, parents, and so on.

During the 2012–13 academic year, we contacted the developers and distributors of the Wilson Reading System and Corrective Reading, and the Torgesen et al. (2006) researchers, inviting them to participate in telephone interviews to answer questions about the program

ingredients. We interviewed the program evaluator who was responsible for overall implementation of the Torgesen et al. (2006) study and who was regularly on-site at the relevant schools during the study period. We interviewed a research executive at Wilson Reading, a Wilson executive who managed program implementation for the study, and the Wilson Language trainer who trained all the teachers who delivered Wilson Reading System during the study. To obtain additional details about program ingredients for Corrective Reading, we interviewed a McGraw-Hill representative responsible for the product, and e-mailed with another. Additionally, we interviewed one of the teachers who delivered Corrective Reading to students throughout the Torgesen et al. study. Interviews ranged in length from 40 minutes to $2^1/_2$ hours. Follow-up questions or clarifications were answered through brief phone calls or via e-mail. Despite the eight-year gap between program implementation and collection of cost data, the verbal reports we elicited on resource requirements were consistent across interviewees except in one instance in which we obtained conflicting reports on the role of a local program coordinator. In this instance we referred to the implementation details reported in Torgesen et al. (2006) and estimated that this personnel requirement would have constituted only around 1% of costs, such that the discrepancy did not substantively affect our analysis.

From these various sources we were able to estimate the number and types of personnel who participated in implementing the two programs, and how much time each person contributed. For Corrective Reading, the main personnel ingredients were the Corrective Reading teachers who worked with students every day. A trained substitute teacher filled in when one of the regular Corrective Reading teachers was absent. Corrective Reading trainers and coaches provided initial professional development and regular ongoing coaching and support to the teachers. The students' regular classroom teachers contributed a very small amount of time to provide information for the selection of students at the outset, and to coordinate logistics for the students throughout the study. Other teachers were involved in initial student screening to assess their reading abilities and need for intervention. A local district coordinator worked across the study schools helping with occasional logistics such as arranging for the substitute teacher. A small amount of parent time was required for attending conferences. Materials required included the Corrective Reading teacher and student materials distributed by McGraw-Hill Education, videos used for training purposes, and screening tests. Other costs incurred included travel for the teachers and trainers for training sessions.

For the Wilson Reading System, the ingredients used in the Torgesen et al. (2006) implementation were very similar to those for Corrective Reading, except that the Wilson Reading System teachers were substantially less experienced (8–9 years of teaching experience vs. 15 years of experience for the Corrective Reading teachers). There was negligible involvement of the classroom teacher and there were no parent conferences. In addition to the teacher and student materials distributed by Wilson Reading, a few classroom materials and supplies were needed such as a magnetic board, binders, markers, and index cards.

Resources devoted to assuring fidelity of implementation, such as having trained observers watch lessons being delivered and providing feedback to the instructors, were included as costs if it appeared that the activities may have affected the impact of the program. Resources that were associated only with the research requirements of conducting an evaluation were not included, as these resources are not relevant for replication of the programs in more general settings, as suggested by Foster, Dodge, and Jones (2003). For example,

administration of posttests was not counted as a program cost if the purpose was simply to determine program impact. Pretests were counted as a cost because they were used as screening measures to determine treatment eligibility or placement.

Associating Costs With Ingredients

Once the type and quantity of each ingredient required to implement each program were specified, the next step was to associate each ingredient with a national price to make the program costs directly comparable. To facilitate the calculation of costs, we listed each ingredient and the quantity required for program implementation in the CBCSE Cost Tool Kit, a set of interlinked Excel spreadsheets designed for this purpose.[4] Item by item, we identified a national average price, mostly from publicly available databases such as the National Occupational Employment and Wage Estimates by the Bureau of Labor Statistics. Details on sources for national prices are available in Hollands et al. (2013). All prices were converted to 2010 dollars for consistency across programs. The program implementations were less than one year in duration so no discounting was necessary except in one sensitivity analysis in which we estimated costs of the programs as implemented over 2–3 years. In this situation we used a conservative 3% discount rate for costs incurred in Years 2 and 3.

In the Torgesen et al. (2006) study, the teachers taught students from both third and fifth grades. We estimated costs attributable to third-grade students to match with the impact data for these same students. For example, the Corrective Reading teacher we interviewed indicated that she spent 33% of her time over the school year teaching three third graders. Reflecting the credentials of the teachers in the study, we found a national average salary for an education specialist with advanced graduate training and 15 years of experience from the Schools and Staffing Survey published by the National Center for Education Statistics (NCES). We calculated 33% of this annual salary and added 31.5% benefits, based on national average benefits rates for public elementary and secondary school employees published by the Bureau of Labor Statistics. We performed similar calculations for all other personnel, using the amounts of time they spent to calculate the fraction of annual salary and benefits that should be attributed to program implementation for the third-grade students. Costs of parent time were calculated using the number of hours contributed and the NCES national average salary for a parent volunteer, to assure that we were considering the opportunity costs of their time. Costs associated with initial training to implement the programs were spread over three years. Although the costs were incurred only in the first year, we assumed that teachers would be able to apply what they learned from the training to their classroom practice for this period of time before needing additional training. We did not spread the costs of ongoing professional coaching and support beyond the year in which they occurred because these needed to be repeated on a regular basis.

Materials and equipment costs were more straightforward to calculate. For example, we obtained the cost of the teacher materials directly from the program distributors and divided the cost of a single set of teacher materials over the number of students for whom they were used to obtain a per-student cost. If the items were durable, we spread the costs over the number of years that our interviewees reported the items were expected to last. Costs of

[4] An online version of the CBCSE Cost Tool Kit is freely available to other researchers for such analyses upon acceptance of a license agreement; see http://www.cbcsecosttoolkit.org/

classroom materials such as a magnetic dry-erase board were obtained by searching online for prices from national distributors. Travel costs for teachers and trainers were based on the amount of time spent traveling (calculated as personnel time as above); costs of transportation, for example, using the car mileage allowance published by the Internal Revenue Service, or the average U.S. domestic itinerary fare for flights published by the Bureau of Transportation Statistics; and, where relevant, hotel and per diem rates published by the General Services Administration.

Facilities costs were more complex to calculate because current market rates such as national average rental rates are not typically available for school buildings. Instead, we used construction costs of school buildings, adjusted for costs of land, development, furnishings, and equipment, and amortized over 30 years. For example, we found a national average construction cost per square foot of an elementary school classroom in the Annual Construction Report published by the School Planning and Management magazine. We uprated this cost per square foot by 33% to account for costs of land, development, furnishings, and equipment (based on *College Planning and Management* magazine, 2011) and amortized the costs over 30 years to obtain the equivalent of a market price per square foot per year for classroom space. Our interviewees gave us information on the size of the classroom they used for pull-out reading instruction or training and the amount of use per year. We arrived at the cost of the classroom space used by each program by multiplying the price per square foot per year by the number of square feet of the classroom, and the fraction of time used per year. We used an interest rate of 3% for amortization, approximating the yield of 30-year U.S. Treasury Bonds. Using a higher interest rate, such as 5%, yielded higher per-student costs for facilities, but because facilities costs were no more than 3%–7% of the total, the relative costs of the programs were not highly sensitive to the interest rate used. A report (Hollands et al., 2013) providing details on ingredients, assumptions, and associated costs was sent to each program developer or distributor and to the evaluator for review.

Cost-Effectiveness Analysis

Cost-effectiveness ratios are calculated by dividing a cost metric, for example, cost per student, by a measure of effectiveness, such as an effect size, in order to demonstrate the cost per unit of improvement in the outcome of interest. In our initial analyses we divided cost per student for each program by the effect sizes reported in Table 1. Because we consider only the costs of the programs above and beyond the resources that students already receive as part of their regular instruction in school, the ratios we report are incremental cost-effectiveness ratios.

Our estimates use average effects among evaluation sites because site-level sample sizes in the Torgesen et al. (2006) study were small and consequently only an overall effect size was calculated. However, in order to accurately assess the resources utilized in attaining specific impacts, we recommend calculating site-level estimates wherever both effectiveness and cost data are obtainable at the site level, especially when implementation varies substantially by site.

Boardman, Greenberg, Vining, and Weimer (2011) suggest three types of sensitivity analysis for cost-effectiveness ratios to assess robustness of the results under different assumptions. The purpose of sensitivity analysis is to explore what the results would be under different assumptions, for example, how the cost-effectiveness ratio might change if programs being compared are implemented with fewer or more students than in the evaluated

implementations, or for a greater or shorter period of time, or using more or less experienced teachers. If the rankings of the programs do not change in these analyses, this suggests that the results are robust. We cannot, however, reliably make assumptions about how changes in implementation will affect impacts; this must be tested empirically. Boardman et al. suggest best- and worst-case sensitivity testing, which places extreme bounds on the results; parameter variation sensitivity testing, where the most influential variables in the model are changed; and Monte Carlo simulation, where the distributions of variables are incorporated into the model. We conducted several sensitivity tests to evaluate the impact of alternate assumptions on our cost-effectiveness ratios and to establish a range of estimates. The major variables and assumptions we tested were: length of program implementation, number of students served by each program or instructor, and the costs of the personnel who represented the most significant expense.

Results

Cost-Effectiveness of Corrective Reading and the Wilson Reading System

Based on the details we obtained regarding the Torgesen et al. (2006) implementation, we estimated costs per student of $6,696 for the Wilson Reading System and $10,108 for Corrective Reading. Tables 2 and 3 list the ingredients, costs per student, and percentage of costs for each program that fell under the categories of personnel, facilities, materials and equipment, and other inputs. As expected for educational programs, the largest percentage of costs was attributable to personnel: over 90% for each program. Further details of program implementation, specific ingredients required, and associated costs are provided in Hollands et al. (2013). Both Corrective Reading and the Wilson Reading System required a significant investment in initial and ongoing training for the instructors, and the instructor salary accounted for the majority of program costs. With a high reliance on fixed-cost components, the cost per student was very dependent on the number of students served per instructor.

Table 2. Ingredients and costs for Corrective Reading per student.

Ingredients	cost per student	% of total costs
Personnel total	$9,542	94%
Corrective Reading teacher	$8,905	
Substitute teacher	$263	
Local district coordinator	$83	
Classroom teacher	$6	
Corrective Reading trainers/coaches	$242	
Testers to screen students	$28	
Parent volunteer time	$14	
Facilities total	$300	3%
Classroom and training facilities[a]	$300	
Materials and equipment total	$135	1%
Lesson materials	$129	
Screening tests	$5	
Training materials	$2	
Other inputs total	$131	1%
Travel costs for Corrective Reading teachers	$23	
Travel cost for trainers	$108	
Grand total	$10,108	100%

[a]Training facilities costs amounted to less than $1 per student. Dollar amounts and percentages may not add exactly to totals shown due to rounding.

Table 3. Ingredients and costs for the Wilson Reading System per student.

Ingredients	Cost per student	% of total costs
Personnel total	$6,063	91%
Wilson Reading System teacher	$5,343	
Substitute teacher	$357	
Local district coordinator	$34	
Wilson Reading System trainers/coaches	$301	
Testers to screen students	$28	
Facilities total	$466	7%
Classroom and training facilities[a]	$466	
Materials and equipment total	$65	1%
Wilson Reading System lesson materials	$50	
Classroom materials	$9	
Screening tests	$5	
Training materials	$1	
Other inputs total	$103	2%
Travel cost for Wilson Reading System teachers	$13	
Travel costs for Wilson Reading System trainers	$89	
Grand total	$6,696	100%

[a]Training facilities costs amounted to less than $1 per student. Dollar amounts and percentages may not add exactly to totals shown due to rounding.

The difference in estimated cost per student between the Wilson Reading System ($6,696) and Corrective Reading ($10,108) can be attributed to several factors, some that are fundamental to the design of each program, and others that are idiosyncratic features of the sample of teachers and students in the Torgesen et al. study. The Corrective Reading teachers had, on average, six extra years of teaching experience compared to the Wilson Reading System teachers, and are therefore assigned higher salaries. Further, the Wilson Reading System teachers taught six third-grade students each on average, compared with an average of only four-and-one-half third graders for each for the Corrective Reading teachers.

Initial cost-effectiveness ratios for the alphabetics domain (i.e., word reading and decoding accuracy and efficiency, as measured by the WRMT letter-word identification, WRMT word attack, and TOWRE sight word efficiency and phonemic decoding efficiency subtests) are presented in Table 4 and suggest that the Wilson Reading System, as implemented in the Torgesen et al. (2006) study, is around twice as cost-effective as Corrective Reading for this

Table 4. Main analysis and sensitivity analyses of cost-effectiveness ratios for the alphabetics domain of Corrective Reading and the Wilson Reading System.

Programs	Total cost per student	Effect size gain	Cost per unit increase in effect size
Main analysis			
Corrective Reading	$10,108	0.22	$45,945
Wilson Reading System	$6,696	0.33	$20,291
Sensitivity analysis 1[a]			
Corrective Reading	$6,332	0.22†	$28,782
Wilson Reading System	$6,188	0.33†	$18,752
Sensitivity analysis 2[b]			
Corrective Reading	$10,784	0.22†	$49,018
Wilson Reading System	$12,674	0.33†	$38,406

† The presumption that the effect sizes observed in the Torgesen et al. (2006) study would hold in the implementations described for the sensitivity analyses should be tested empirically. Sensitivity analyses should be interpreted only as "what-if" scenarios. [a]Assumes teachers with five years of experience teaching four groups of three students. [b]Assumes full implementation of program over two to three years.

domain: $20,291 per standard deviation increase in effect size for the Wilson Reading System versus $45,945 for Corrective Reading. This reflects substantially lower costs for the Wilson Reading System and greater effectiveness for alphabetics. In a sensitivity analysis, we found that if both programs were delivered in the study by teachers with five years of teaching experience to four groups of three students each, the costs per student would be very close: $6,188 for the Wilson Reading System and $6,332 for Corrective Reading. If the same effect sizes observed for alphabetics by Torgesen et al. applied in this scenario, the cost-effectiveness ratios would improve to $18,752 per standard deviation increase in alphabetics skills for the Wilson Reading System and $28,782 for Corrective Reading.

In addition to improving alphabetics, Corrective Reading showed a positive impact on text reading fluency, as measured by the Edformation Oral Fluency Assessment (effect size 0.27), yielding a cost-effectiveness ratio of $37,437 per standard deviation increase in fluency. For the Wilson Reading System, impact on fluency (effect size 0.15) is not statistically significant. In the absence of evidence of effectiveness in this domain, we refrain from presenting a cost-effectiveness ratio for the Wilson Reading System for fluency (which could technically reach infinity if the effect size was 0). We conclude that Corrective Reading, at least in the Torgesen et al. (2006) implementation, is more cost-effective for fluency than the Wilson Reading System. If choosing between the two programs, decision makers should consider the relative importance of addressing alphabetics and fluency for their particular student populations to help assess whether Corrective Reading's greater cost-effectiveness with respect to fluency offsets its lower cost-effectiveness with respect to alphabetics. In their modified versions, neither program had an impact on comprehension but we expect that full program implementations would yield different results for both costs and other outcomes such as vocabulary or comprehension.

We conducted a second sensitivity analysis to model the costs of implementing the two programs under typical school circumstances and for the full dosage and duration recommended by the developers (240 hours of instruction for Corrective Reading in 45-minute periods four to five times per week over two to three years, and 60–90 minute periods three to five times per week over two to three years for the Wilson Reading System). We gathered additional information from the developers and distributors of the programs to establish details of resource requirements over the longer period of time. In a recommended implementation of the Wilson Reading System, a student would receive around 450 hours of instruction, compared with 240 for Corrective Reading. Teachers of the Wilson Reading System must be certified, which requires, at minimum, participation in a three-day introductory workshop every five years, 90 hours of online professional development, 60 hours of practicum, and five observations per year from a certified Wilson Reading trainer. Teachers of Corrective Reading typically participate in one day of initial training and one day per year of coaching. For these analyses we assumed that each Corrective Reading or Wilson Reading System teacher taught four groups of four students for the program duration, in line with developer recommendations.

We found that Corrective Reading is only around 7% more costly in a typical implementation over two years ($10,784 vs. $10,108 per student) than in the one-year Torgesen et al. (2006) implementation. Efficiency is clearly improved by serving 16 students per teacher rather than 9–12 per teacher as in the Torgesen et al. study. Training requirements in practice are minimal, while in the study teacher training and ongoing support were very intensive and therefore expensive. The Wilson Reading System is about twice as costly over $2^1/_2$ years

as it was in the Torgesen et al. year-long implementation: $12,674 per student vs. $6,696. This reflects the fact that, in addition to 360 more hours of instruction per student, the training requirements for Wilson Reading System teachers in the field are more aligned with those that occurred in the study. Some efficiency is gained by serving 16 students per teacher rather than 12 per teacher as in the Torgesen et al. study.

We were not able to find rigorous studies of Corrective Reading or the Wilson Reading System in which similar populations of students were exposed to the program for the full length of time recommended by the program developers. In the absence of relevant effect sizes, we inferred that, because a full implementation of each program would include the vocabulary and comprehension components, the impact on alphabetics might be similar to that observed in Torgesen et al. (2006), but that improvements might be seen in fluency, vocabulary, and comprehension outcomes. Using our cost estimates for the full-length program implementations and the Torgesen et al. (2006) alphabetics effect sizes, we show cost-effectiveness ratios resulting from this analysis in Table 4. This represents a "what-if" scenario that could only be confirmed through a rigorous, long-term study of the programs. Despite almost doubling in costs, the Wilson Reading System remains more cost-effective for alphabetics than Corrective Reading, indicating that our initial conclusion with respect to cost-effectiveness ranking is robust.

Discussion and Implications for Research on the Cost-effectiveness of Reading Programs

This study was intended to demonstrate the application of cost-effectiveness analysis to early reading programs. We found that the lack of comparability in the available effectiveness data on multiple reading outcomes presents challenges for comparing potential alternative programs with respect to both effectiveness and cost-effectiveness. We estimated costs for one pair of programs with comparable effectiveness estimates using a uniform method of gathering cost data and combined these cost data with the available effectiveness data to illustrate the potential value of cost-effectiveness ratios for decision makers choosing among alternative programs. In our comparison, we found that even for similar programs that show evidence of effectiveness, the cost-effectiveness ratios can differ dramatically, both due to differences in program effectiveness and to substantial differences in costs of implementation.

Through sensitivity analyses in which we varied the amount and costs of ingredients, we were able to explore how total program costs and cost-effectiveness ratios might deviate from our initial estimates under different scenarios, assuming that the observed impacts hold under these altered circumstances. This assumption regarding impacts may be reasonable in situations where the change is unlikely to affect the quality of program delivery, for example, if a teacher with 15 years of experience is substituted for one having 20 years. However, the assumption may not be feasible if the teacher is replaced with a novice or a parent volunteer. We show how decision makers can use sensitivity analyses to explore the impact of changes in implementation on costs per student and the cost-effectiveness ratios. However, we stress that these are only demonstrative and that the impact of such changes on reading outcomes would need to be tested empirically.

Beyond the execution of sensitivity analyses, greater confidence in the cost-effectiveness results for a program could be established by calculating cost-effectiveness ratios for a

number of different studies and implementations of the same program, in order to present a range. Decision makers could use these results to help choose among viable alternative programs while carefully considering how variations in program implementation made to accommodate budget pressures or to fit in with existing schedules, curriculum, and personnel availability might affect the expected impact of a program, as well as the costs.

Many of the methodological challenges involved in performing cost-effectiveness analysis of early reading instruction could be alleviated by changes in how the effectiveness of reading interventions is evaluated and perhaps in changes in the incentives offered to researchers evaluating reading interventions (e.g., from funding organizations and academic norms). Our application of cost-effectiveness analysis, which is rarely performed in education, depends critically on having comparable effectiveness outcomes. Differences in age and reading level as well as ethnicity and socioeconomic status of populations served in the evaluations of reading programs, and in measures used to assess impact, present a challenge not only for comparisons of program effectiveness and for cost-effectiveness analysis, but also for policymakers who need to decide which intervention to select. We suspect that the lack of comparable effectiveness evidence across studies is due in part to incentives that have encouraged researchers to develop their own individual interventions and demonstrate that they work in comparison to business-as-usual, rather than in comparison to alternative interventions. Such incentives may include the priorities of funding organizations but also norms and expectations in higher education that encourage researchers to establish their own programs as effective. Studies that compare the impact of reading programs with each other, rather than simply investigating the impact of one program compared with business-as-usual, would be helpful to decision makers in selecting programs that might reasonably serve as alternative choices. We additionally recommend that all rigorous program evaluations should collect data on costs using the ingredients method.

An unresolved methodological and conceptual challenge for valuing early reading programs lies in how to aggregate effectiveness data across multiple related outcomes for programs that target a variety of proximal and distal reading skills. There is no obvious way to combine the results across multiple outcome domains (alphabetics, fluency, comprehension, and general reading achievement in this analysis) for the purposes of estimating a single summary of a program's cost-effectiveness. As noted by Levin (1975) and by Levin and McEwan (2001), this challenge arises in valuing many educational programs because they often have an impact on more than one learning outcome. Although cost-effectiveness analysis allows a decision maker to consider each outcome individually (e.g., based on students' instructional needs) and select the program that is most cost-effective in addressing that outcome, this approach does not provide a single metric that reflects the overall cost-effectiveness across multiple outcomes. Future research could explore the application of multiattribute utility theory (Edwards & Newman, 1982; Levin & McEwan, 2001) to reading programs that affect multiple reading outcomes, recognizing that such an analysis would be complicated by the fact that reading outcomes are not independent of each other, but rather developmentally dependent in complex ways.

A second solution is to compare the cost-effectiveness of reading programs with respect to reading comprehension, the ultimate goal of any reading intervention, even when programs target more proximal outcomes such as alphabetics. In our analysis, such a comparison was not viable because Corrective Reading and the Wilson Reading System were modified in the Torgesen et al. (2006) implementation to exclude the comprehension

components, such that measured impacts on reading comprehension showed nonsignificant results. Our focus on alphabetics, while appropriate given the existing data, should not be interpreted as capturing what is most important among reading outcomes. We recommend that future evaluations of early reading programs include comparable measures of reading comprehension and, ideally, some comparable measures of each of the targeted proximal and distal reading outcome domains. Common measures would also eliminate the need to calculate effect sizes, which are hard to interpret and tend to be larger when the study population is homogeneous (Fern & Monroe, 1996; Olejnik & Algina, 2000).

For future evaluations of early reading programs, we recommend the design and inclusion of cost analyses simultaneously with determinations of program effectiveness in order to facilitate the most accurate and timely assessment of cost-effectiveness. WWC currently sets clear, rigorous standards for what constitutes a credible impact evaluation and could establish similar standards for the concurrent collection and analysis of cost data in a standardized manner that would facilitate comparability across programs. Levin and Belfield (2015) set out a number of recommendations with respect to the collection of cost and cost-effectiveness data that could serve as a starting point for such standards. These include a suggestion that, wherever feasible, programs should be evaluated at multiple sites with large-enough sample sizes to allow an investigation of how effectiveness and resource use vary across sites. Inclusion of qualitative descriptions of implementation to explain the challenges and advantages encountered at particular sites would facilitate an assessment of how resource use and resource management are related to effectiveness. In situations where the cost-effectiveness of a program varies substantially across sites, this qualitative information may provide insights to explain the differences in efficiency of resource use and to identify the most productive sites. It would also allow decision makers to select programs that have been shown to be effective under conditions similar to those in which they are operating.

Cost-effectiveness evaluations of early reading programs that are widely used in American schools would provide valuable information to decision makers who need to make choices from a realistic menu of options when making resource allocation decisions. Although decisions regarding reading programs must involve a variety of contextual considerations including the nature of the student population to be served and local stakeholder preferences, cost-effectiveness data can assist in increasing the efficiency of resource use. Since the creation of the Institute of Education Sciences within the U.S. Department of Education, federal funding has led to a proliferation of rigorous evaluation studies. Although this has improved the availability of effectiveness data, questions remain about how relevant and useful those data are to education decision makers. We argue that government agencies and private foundations could encourage the collection and use of rigorous and comparative cost data in combination with effectiveness data in making decisions regarding program funding, a recommendation that we would have likely also made at the onset of this study. An additional recommendation that emerged from this study is that funding organizations should also encourage the accumulation of more comparable effectiveness data—by incentivizing researchers to contrast alternative programs with one another, when possible, and to conduct studies that align more closely with existing studies in their outcome domains, measures, and populations when comparing an individual intervention to business-as-usual. By improving the efficiency of resource allocation among programs, education decision makers can improve the productivity of education.

Funding

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Award Number R305U120001 to Teachers College, Columbia University.

ARTICLE HISTORY

Received 18 July 2014
Revised 18 May 2015
Accepted 23 May 2015

EDITORS

This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2011). *Cost-benefit analysis: Concepts and practice* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis*, *24*(4), 243–266.

Bowman-Perrott, L. J., Herrera, S., & Murry, K. (2010). Reading difficulties and grade retention: What's the connection for English language learners? *Reading and Writing Quarterly*, *26*(1), 91–107.

Chambers, J., & Parrish, T. (1994a). Developing a resource cost database. In W. S. Barnett (Ed.), *Cost analysis for education decisions: Methods and examples* (Vol. 4, pp. 23–44). Greenwich, CT: JAI.

Chambers, J., & Parrish, T. (1994b). Modeling resource costs. In W. S. Barnett (Ed.), *Cost analysis for education decisions: Methods and examples* (Vol. 4, pp. 7–21). Greenwich, CT: JAI.

Chambers, J. G., Parrish, T. B., & Harr, J. J. (2004). *What are we spending on special education services in the United States, 1999–2000*? Special Education Expenditure Project (SEEP). Retrieved from http://csef.air.org/publications/seep/national/advrpt1.pdf

College Planning and Management. (2011, June). *Living on campus. Trends and analysis*. Retrieved from https://webcpm.com/research/2011/06/college-housing/asset.aspx.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., … Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass Publishers.

Duke, N. K., & Block, M. K. (2012). Improving reading in the primary grades. *Future of Children*, *22*(2), 55–72.

Edwards, W., & Newman, J. R. (1982). *Multiattribute evaluation*. Thousand Oaks, CA: Sage Publications.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393–447.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta analysis. *Reading Research Quarterly*, *36*(3), 250–287.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1–44.

Every Child a Chance Trust. (2009). *The long term costs of literacy difficulties* (2nd ed.). Retrieved from http://readingrecovery.org/images/pdfs/Reading_Recovery/Research_and_Evaluation/long_term_costs_of_literacy_difficulties_2nd_edition_2009.pdf

Fern, F. E., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*(2), 89–105.
Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*, 37–55.
Foster, E., Dodge, K. A., & Jones, D. (2003). Issues in the economic evaluation of prevention programs. *Applied Developmental Science*, *7*(2), 76–86.
Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*(1), 3–17.
Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First impact study final report* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3–29.
Hatcher, P. J., Goetz, K., Snowling, M. J., Hulme, C., Gibbs, S., & Smith, S. (2006). Evidence for the effectiveness of the Early Literacy Support programme. *British Journal of Educational Psychology*, *76*, 351–367.
Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). Empirical benchmarks for interpreting effect sizes in research *(MDRC Working Paper)*. Retrieved from http://www.mdrc.org/sites/default/files/full_84.pdf
Hollands, F., Bowden, A. B., Belfield, C., Levin, H.M., Cheng, H., Shand, R., … Hanisch-Cerda, B. (2014). Cost-effectiveness analysis in practice: Interventions to improve high school completion. *Educational Evaluation and Policy Analysis 36*(3), 307–326.
Hollands, F. M., Pan, Y., Shand, R., Cheng, H., Levin, H.M., Belfield, C. R., … Hanisch-Cerda, B. (2013). *Improving early literacy: Cost-effectiveness analysis of effective reading programs*. Retrieved from http://cbcse.org/wordpress/wp-content/uploads/2013/05/2013-Hollands-Improving-early-literacy1.pdf
Hummel-Rossi, B., & Ashdown, J. (2010). *Cost-effectiveness analysis as a decision tool in selecting and implementing instructional interventions in literacy*. Retrieved from http://fdf.readingrecovery.org/component/content/article/154-cost-effectiveness-analysis
Jimerson, S. R., & Kaufman, A. M. (2003). Reading, writing, and retention: A primer on grade retention research. *The Reading Teacher*, *56*(7), 622–635.
Justice, L., Kaderavek, J., Fan, X., Sofka, A., & Hunt, A. (2009). Accelerating pre-schoolers' early literacy development through classroom-based teacher-child storybook reading and explicit print referencing. *Language, Speech and Hearing Services in Schools*, *40*, 67–85.
Kieffer, M. J. (2011). Converging trajectories: Reading growth in language minority learners and their classmates, Kindergarten to Grade 8. *American Educational Research Journal*, *48*(5), 1187–1225.
Kim, J. S., Olson, C. B., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D., … Land, R. E. (2011). A randomized experiment of a cognitive strategies approach to text-based analytical writing for mainstreamed Latino English language learners in Grades 6–12. *Journal of Research on Educational Effectiveness*, *4*(3), 231–263.
Levin, H. M. (1970). A cost-effectiveness analysis of teacher selection. *The Journal of Human Resources*, *5*(1), 24–33.
Levin, H. M. (1975). Cost-effectiveness in evaluation research. In M. Guttentag & E. Struening (Eds.), *Handbook of educational research* (pp. 89–122). Thousand Oaks, CA: Sage Publications.
Levin, H. M. (2001). Waiting for Godot: Cost-effectiveness analysis in education. *New Directions for Evaluation*, *2001*(90), 55–68.
Levin, H. M. (2011). The consideration of costs in improving literacy. In D. Lapp & D. Fisher (Eds.), *Handbook of research on teaching the English language arts* (3rd ed., pp. 120–124). New York, NY: Routledge.
Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*, *8*, 400–418. doi:10.1080/19345747.2014.915604

Levin, H. M., Belfield, C. R., Hollands, F. M., Bowden, A. B., Cheng, H, Shand, R., … Hanisch-Cerda, B. (2012). *Cost-effectiveness analysis of interventions that improve high school completion*. Retrieved from http://cbcse.org/wordpress/wp-content/uploads/2012/10/HighSchoolCompletion_REVISED.pdf

Levin, H. M., Catlin, D., & Elson, A. (2007). Costs of implementing adolescent literacy programs. In D. Deshler, A. S. Palincsar, G. Biancarosa, & M. Nair (Eds.), *Informed choices for struggling adolescent readers: A research-based guide to instructional programs and practices* (pp. 61–91). Newark, DE: International Reading Association.

Levin, H. M., Glass, G. V., & Meister, G. R. (1987). Cost-effectiveness of computer-assisted instruction. *Evaluation Review*, *11*(1), 50–72. doi: 10.1177/0193841×8701100103

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Lonigan, C. J., & Shanahan, T. (2009). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

Massoni, S., & Vergnaud, J. (2012). How to improve pupils' literacy: A cost-effectiveness analysis of a French educational project. *Economics of Education Review*, *31*(1), 84–91.

McArthur, G. (2008). Does What Works Clearinghouse work? A brief review of Fast ForWord. *Australasian Journal of Special Education*, *32*(1), 101–107.

National Center for Education Statistics. (2014). *The Condition of Education 2014*. Retrieved from http://nces.ed.gov/pubs2014/2014083.pdf

National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Nelson, J. R., Benner, G. J., & Gonzales, J. (2005). An investigation of the effects of a prereading intervention on the early literacy skills of children at risk of emotional disturbance and reading problems. *Journal of Emotional and Behavioral Disorders*, *13*(1), 3–12.

Nelson, J. R., Stage, S. A., Epstein, M. H., & Pierce, C. D. (2005). Effects of a prereading intervention on the literacy and social skills of children. *Exceptional Children*, *72*(1), 29–45.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology*, *1*, 435–450.

Pikulski, J. (1994). Preventing reading failure: A review of five effective programs. *The Reading Teacher*, *48*(1), 30–39.

Ritchey, K., & Goeke, J. (2006). Orton-Gillingham and Orton-Gillingham-based reading instruction: A review of the literature. *The Journal of Special Education*, *40*(3), 171–183.

Roberts, G., Mohammed, S.S., & Vaughn, S. (2010). Reading achievement across three language groups: Growth estimates for overall reading and reading subskills obtained with the Early Childhood Longitudinal Survey. *Journal of Educational Psychology*, *102*(3), 668–686.

Rouse, C. E., & Krueger, A.B. (2004). Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program. *Economics of Education Review*, *23*, 323–338.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literature research* (pp. 97–110). New York, NY: Guilford Press.

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the Reading Recovery early intervention. *Journal of Educational Psychology*, *97*(2), 257–267.

Scientific Learning Corporation. (2005). Improved early reading skills by students in three districts who used Fast ForWord to Reading 1. *MAPS for Learning: Product Reports*, *9*(1), 1–5.

Simon, J. (2011). *A cost-effectiveness analysis of early literacy interventions* (Doctoral dissertation). Columbia University, New York, NY.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14.

Slavin, R., & Madden, N. A. (2011) Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370–380. doi: 10.1080/19345747.2011.558986

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

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*(1), 72–110.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., … Compton, D. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across sites and years. *Educational Evaluation and Policy Analysis*, *30*(4), 368–388.

Stockard, J. (2008). *The What Works Clearinghouse Beginning Reading reports and rating of Reading Mastery: An evaluation and comment* (NIFDI Technical Report 2008-4). Eugene, OR: National Institute for Direct Instruction.

Stockard, J., & Wood, T. W. (2013). *The WWC review process: An analysis of errors in two recent reports* (NIFDI Technical Report 2013-4). Eugene, OR: National Institute for Direct Instruction.

Torgesen, J. K., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., … Haan, C. (2006). *National Assessment of Title I: Interim Report. Volume II. Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers*. Retrieved from http://www2.ed.gov/rschstat/eval/disadv/title1interimreport/vol2.pdf

Tsang, M. C. (1997). Cost analysis for improved policy-making in education. *Educational Evaluation and Policy Analysis*, *19*(4), 18–24.

U.S. Department of Education. (1997). *Time spent teaching core academic subjects in elementary schools: Comparisons across community, school, teacher and student characteristics*. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97293

U.S. Government. (2013). *U.S. government spending*. Retrieved from http://www.usgovernmentspending.com/year_spending_2013USbn_13bs2n_20#usgs302

Vadasy, P. F., & Sanders, E.A. (2008). Code-oriented instruction for kindergarten students at risk for reading difficulties: A replication and comparison of instructional groupings. *Reading and Writing: An Interdisciplinary Journal*, *21*(9), 929–963.

WWC. (2007a). *WWC intervention report: Corrective Reading*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/WWC_Corrective_Reading_070207.pdf

WWC. (2007b). *WWC intervention report: Wilson Reading System*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/WWC_Wilson_Reading_070207.pdf

WWC. (2012). *WWC evidence review protocol for beginning reading interventions* (Version 2.1). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/br_protocol_v2.1.pdf

WWC. (2013). *What Works Clearinghouse: Procedures and standards handbook* (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/refereny ce_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf