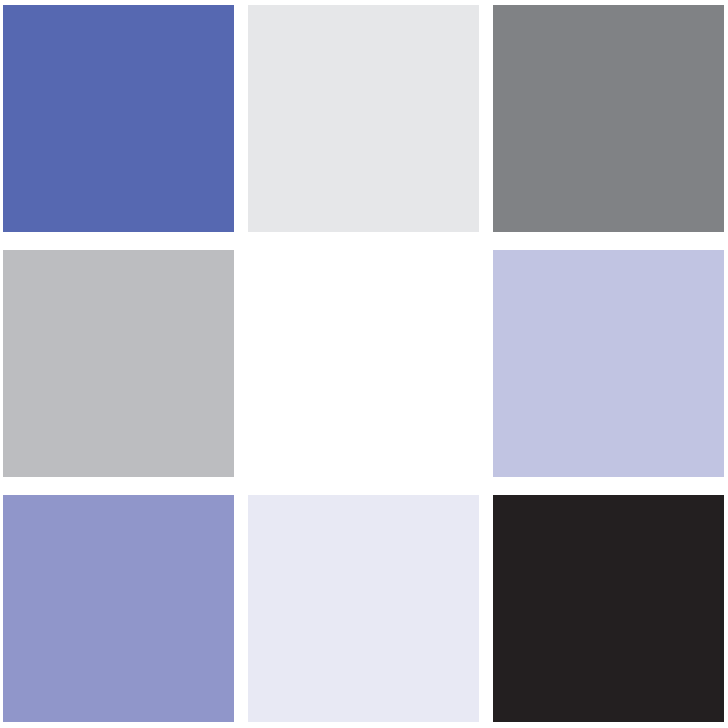
IMPROVING EARLY LITERACY: COST-EFFECTIVENESS ANALYSIS OF EFFECTIVE READING PROGRAMS

April 2013



Fiona M. Hollands

Yilin Pan

Robert Shand

Henan Cheng

Henry M. Levin

Clive R. Belfield

Michael Kieffer

A. Brooks Bowden

Barbara Hanisch-Cerda

Center for Benefit-Cost Studies of Education

Teachers College, Columbia University

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Award Number R305U120001 to Teachers College, Columbia University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

CONTENTS

ACKNOWLEDGEMENTS

Our ability to obtain accurate and detailed information regarding the ingredients required to implement the early literacy programs included in this study depended almost entirely on the willingness and patience of the relevant program evaluators, developers, and implementers to entertain many detailed questions. While we respect the confidentiality of our interviewees by not naming them here, we are deeply grateful for their time.

SUMMARY

This paper calculates the cost-effectiveness of seven reading programs that have been shown to be effective with respect to early literacy outcomes for students in kindergarten through third grade. Three programs serve kindergarten students: Kindergarten Peer-Assisted Learning Strategies (K-PALS), Stepping Stones to Literacy (Stepping Stones), and Sound Partners. Two of the programs primarily serve first graders: Fast ForWord Reading 1 (FFW1) and Reading Recovery. Two others serve third grade students: Corrective Reading and Wilson Reading System. All programs serve below-average or struggling readers by providing instruction that is supplementary to regular classroom instruction, except for K-PALS, which serves all readers in the regular classroom and is a partial substitute for classroom reading instruction.

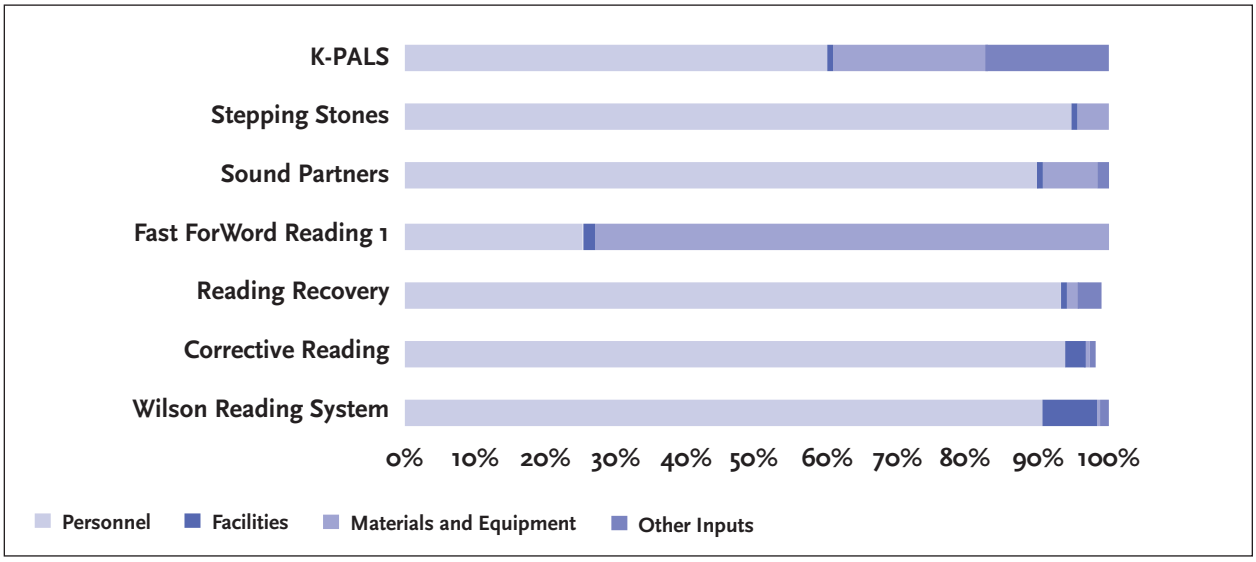
Effectiveness of each program in improving outcomes in alphabetics, fluency, and reading comprehension (see Appendix I for definitions) was obtained from the What Works Clearinghouse. All seven programs showed positive impact on at least one measure of alphabetics. Three programs were also effective at improving reading fluency, and one program showed an additional positive impact on reading comprehension.

Cost data for each program were collected using the ingredients method. Program evaluators, developers and implementers were interviewed to obtain detailed information regarding the resources required to implement each program as it was evaluated. We focused on incremental costs of delivery, i.e., costs above and beyond what was already being spent on regular school programming. In one case where we could not obtain ingredients for the evaluated implementation, we costed out the ingredients required for an average implementation of the program. While our preference is to match site-level costs to site-level effectiveness data, we were only able to do this for one program as most evaluations involved small numbers of students at each site, precluding site-level impact analysis. As shown in Figure S1, personnel accounted for the most significant portion of program costs in all cases but one, FFW1, which is the only computer-based program of the seven.

Total program costs were spread over the number of students participating in the program at the study sites in order to obtain costs per student for each program. Costs per student across programs generally increased substantially with the grade of the students being served, perhaps reflecting the increasing seriousness of reading problems or the difficulty rectifying issues that were not addressed in earlier grades. Other factors affecting costs included whether the program substituted existing instruction or supplemented it; how long the intervention lasted; and whether the instructors were tutors or specially trained teachers. The range of costs was approximately $30 per student for K-PALS to over $10,000 per student for Corrective Reading.

Figure S1

Distribution of Program Costs Across Major Ingredients Categories



Incremental costs per student for each program were combined with effect sizes to obtain incremental cost-effectiveness ratios. For programs that showed a positive impact on more than one literacy outcome, we split the costs across the outcomes based on percent of program delivery time spent addressing each outcome, as reported by the program developers. We found very large differences in the cost-effectiveness of the seven programs, as summarized in Table S1. For the alphabetics domain, the incremental cost-effectiveness ratios to obtain a unit increase in effect size ranged from a low of $38 for K-PALS to a high of $38,135 for Corrective Reading. For the fluency domain, the incremental cost-effectiveness ratios to obtain a unit increase in effect size ranged from a low of $165 for Sound Partners to a high of $6,364 for Corrective Reading. For each program, we conducted one or more sensitivity tests to assess the impact of different assumptions on the cost-effectiveness ratios. The majority of these involved varying the most important ingredients used in the implementation, their costs, or the number of students being served by each instructor.

The goal of cost-effectiveness analysis is to compare alternative programs for efficiency of resource use. Differences in age and reading ability of the students targeted by the seven programs limited the number of relevant comparisons. We present comparisons only among programs serving the same grade level and reading ability. At the kindergarten level, we find that Stepping Stones is more cost-effective in the alphabetics domain than Sound Partners, but Sound Partners also has a positive impact on both fluency and reading comprehension. At the first grade level, we find that FFW1 is more cost-effective than Reading Recovery for the alphabetics domain, but Reading Recovery also has a positive impact on fluency. Finally, at the third grade level, Wilson Reading System appears to be more cost-effective than Corrective Reading for the alphabetics domain, but Corrective Reading has an additional positive impact on fluency.

One issue that remains unresolved is how to value the programs to account for impact on multiple literacy domains. We explore various alternatives but conclude that there is no satisfactory objective solution beyond simply comparing program impact on the ultimate goal of literacy programs: reading comprehension. However, an individual decision-maker can assign subjective weights to the cost-effectiveness

ratios for different domains based on his/her knowledge of the literacy needs of the student population being served.

We recommend that future evaluations of reading programs include common outcome measures to facilitate comparability among programs. Studies in which two or more alternative programs are implemented with similar populations of students, and literacy outcomes are compared using the same measures, would greatly facilitate comparability not only of program effectiveness, but also of cost-effectiveness. We also suggest that cost data should be collected concurrently with effectiveness data to allow the most accurate documentation of resource requirements. Armed with data on both costs and effects of alternative literacy programs, education decision-makers can include program efficiency among the decision criteria used to select a specific program for implementation.

Table S1

Summary Characteristics and Cost-effectiveness Ratios of Effective Early Literacy Programs

Programs by grade level	Reading ability of target students	Program duration (weeks)	Total cost per student	Literacy domain	Effect size gain	Cost per unit increase in effect size*
Kindergarten average readers:						
K-PALS**	All	20	$27	Alphabetics	0.61	$38
Kindergarten struggling readers:						
Stepping Stones	Struggling; behavioral disorders	5	$479	Alphabetics	0.84	$570
Sound Partners	20–30th percentile	18	$791	Alphabetics	0.34	$2,093
				Fluency	0.48	$165
First grade struggling readers:						
Fast ForWord Reading 1	Slightly below average	6	$282	Alphabetics	0.24	$601
Reading Recovery	Bottom 20th percentile	12–20	$4,144	Alphabetics Fluency	0.70 1.71	$1,480 $606
Third grade struggling readers:						
Corrective Reading	Bottom 25th percentile	28	$10,108	Alphabetics Fluency	0.22 0.27	$38,135 $6,364
Wilson Reading System	Bottom 25th percentile	28	$6,696	Alphabetics	0.33	$13,392

* Note that the cost per student is adjusted by the amount of program delivery time that addresses each literacy domain in order to calculate the cost-effectiveness ratio.

** Workshop level of implementation.

1. INTRODUCTION

Almost 40% of the elementary school day is devoted to the subjects of English, reading, and language arts, all contributing towards the development of literacy (USDOE, 1997). By comparison, only 16% of the day is spent on mathematics and about 9% each for science and social studies. With average total expenditures per student in U.S. public schools at $12,643 in 2008-2009 (NCES, 2012), spending on literacy is approximately $5,000 per student per year. The results of this substantial investment are mixed, with 33% of students in Grade 4 scoring below a basic level of proficiency in reading, as measured by National Assessment of Educational Proficiency (NAEP) tests.[1] Clearly there is a need to identify and implement literacy interventions that are effective for a greater number of elementary school students, particularly struggling readers who may not be well served by the existing reading curricula in their schools.

Early investment in development of reading skills and remediation of reading difficulties is critical because early literacy is significantly associated with later academic achievement (Duncan et al., 2007; Hernandez, 2012). Indeed, Snow, Burns, and Griffin (1998) assert that many reading problems experienced by adolescents and adults arise from issues that could have been addressed in early childhood. They stress the importance of helping children overcome literacy obstacles in the primary grades or earlier. According to the Committee on the Prevention of Reading Difficulties in Young Children, all primary-grade classrooms should attend to "the alphabetic principle, reading sight words, reading words by mapping speech sounds to parts of words, achieving fluency, and comprehension" (Snow et al., 1998, p.6). This Committee recommends that effective support for struggling readers involve supplementary instruction delivered by a trained reading specialist to individual students or small groups, in close coordination with high quality instruction from the classroom teacher. Furthermore, regular communication between the elementary classroom teachers and reading specialists within a school is required to facilitate identification of students' literacy needs, and both sets of professionals should benefit from ongoing professional development and collegial support. There are many programs and interventions that satisfy these general requirements. Schools and districts therefore face important decisions in identifying which specific programs are most effective.

Early literacy programs selected for implementation should not just be effective; given the amounts spent, these programs should also be the most cost-effective. Relative costs of the programs should be included in the criteria used in making a decision about which program to adopt. Levin (2001, 2011), Harris (2009), and others, have argued for the importance in considering both effectiveness data and costs when choosing among several alternative interventions targeting the same outcome. Tsang (1997) suggests that cost-effectiveness analysis of alternative educational interventions can inform educational reform and lead to substantial cost savings. Ross, Barkaoui, and Scott (2007) offer a similar argument that information on costs can be used to rewrite regulations and improve the efficiency of programs. Studies of cost-effectiveness have been conducted on a range of education-related topics such as teacher selection and training, educational technology, math curricula, increasing the length of the school day, peer tutoring and reduction in class size (Levin, 1995). These studies help decision-makers choose between effective programs that differ in the resources required to implement them.

Few empirical studies have been conducted on the cost-effectiveness of literacy programs (Hummel-Rossi & Ashdown, 2010). One notable exception is Simon's (2011) cost-effectiveness analysis of four early

1 http://nces.ed.gov/programs/coe/tables/table-rd2-1.asp.

literacy programs: Classwide Peer Tutoring, Reading Recovery, Success for All, and Accelerated Reader. Combining effectiveness data with cost data collected retrospectively, Simon (2011) found significant differences across the four programs in costs ($500-$11,700 per student per year), effects, and cost-effectiveness ($1,400-$45,000 per unit increase in effect size for literacy outcomes). This evidence suggests a strong possibility that resources deployed to improve early literacy may be allocated more efficiently. We build on Simon's work by tying the costs of program implementations to the effects on literacy observed as a result of those specific implementations. We are also able to investigate programs serving students in specific grade levels in elementary school and to compare programs serving the same grade.

Cost analysis by Levin, Catlin, and Elson (2007) of three adolescent reading programs also illustrates substantial variation in costs across programs and extends the analysis to site-specific variation. Costs of implementing READ 180, Questioning the Author, and Reading Apprenticeship varied significantly among sites even for the same program. For example, the cost of implementing READ 180 was as low as $285 per student at one site and as high as $1,510 per student at another site. If the program were implemented according to the developer's recommendations, the costs should be approximately $1,100. Such differences illustrate the importance of costing out programs based on actual site-level implementation data.

In this paper, we focus on seven interventions that have been demonstrated to improve early literacy. We apply the ingredients method to calculate costs and subsequently derive cost-effectiveness ratios for the interventions. Ideally, cost-effectiveness analysis should allow us to compare the interventions and make recommendations to decision-makers as to which interventions are preferable based on economic reasoning (i.e., not accounting for politics or feasibility). However, our analysis suggests caution in making such comparisons and policy recommendations across the seven programs. As we show below, differences among the interventions and how they were studied in grade-level and reading ability of the students targeted preclude a direct comparison across all seven programs. They are not interchangeable alternatives addressing the exact same problem. Instead, we present comparisons only among programs addressing students of similar reading ability and in the same school grade. We also aim to highlight the key methodological and empirical challenges in performing cost-effectiveness analysis of literacy programs, and highlight important gaps in the existing research base, with the intention to improve future research practice.

Our analysis proceeds as follows: Section 2 describes early literacy outcomes, our selection process for the interventions included in this study, the effectiveness data we use, our methods for collecting cost data, and the comparability of these interventions; Section 3 presents the cost analysis and cost-effectiveness results for each program; Section 4 provides comparisons, summary, and discussion; and Section 5 offers some conclusions and suggestions for further research. Appendices provide definitions of terms used in the paper, the protocol used to gather evidence on ingredients used to implement each intervention, sources of national prices, and abbreviations.

2. EARLY LITERACY: MEASURING THE EFFECTIVENESS AND COSTS OF INTERVENTIONS

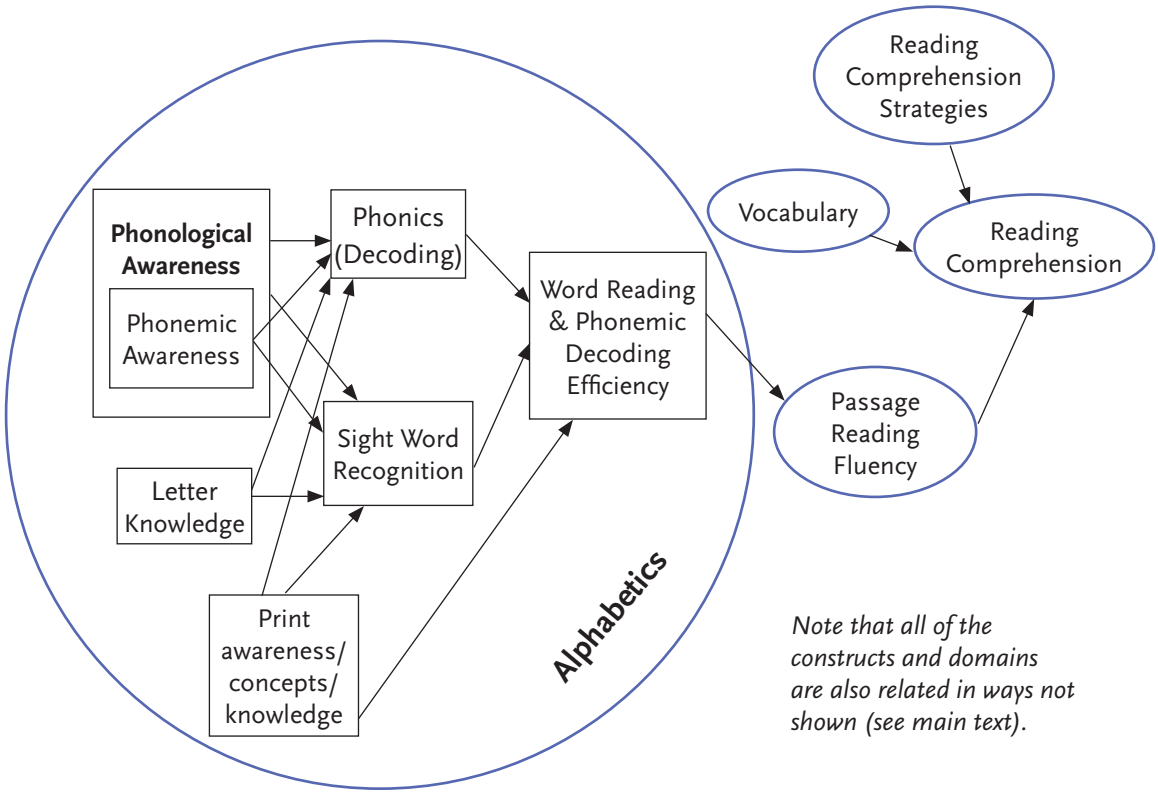
2.1 Defining the Outcomes of Early Literacy Interventions

The first task in performing cost-effectiveness analysis is to identify interventions that target the same outcome and have comparable measures of effectiveness. Unfortunately, there is some disagreement about how to classify early literacy outcomes. Distinguishing literacy outcomes for our purposes is particularly challenging because these outcomes are hierarchically and causally related (e.g., Snow, Burns, & Griffin, 1998). The 2000 Report of the National Reading Panel (NICHD, 2000) defines three overarching categories of outcomes: alphabetics, reading fluency, and comprehension. But these categories are not independent of each other and are more appropriately regarded as sequential.

Figure 1 provides a heuristic of how these categories might be considered to relate to each other, and illustrates more specific outcomes within the broad categories. On this heuristic, literacy domains are represented as circles and literacy constructs within domains are represented as squares. This heuristic is intended not as a complete model, but as a useful simplification for our purposes, given that many of the relations displayed as unidirectional are in fact considered reciprocal (e.g., vocabulary and reading comprehension;

Figure 1

Heuristic for the relations among proximal and distal reading domains, as informed by NICHD (2000) and the WWC classifications, with domains represented as circles and constructs within domains represented as squares.



Stanovich, 1986), many of the precursor skills are developmentally related as well (e.g., vocabulary and phonological awareness; Metsala & Walley, 1998), and that relations differ across development.

The What Works Clearinghouse (WWC), a national database of research reviews on the effectiveness of rigorously evaluated interventions, identifies 35 early (K-3) literacy programs that have positive or potentially positive effects for one or more of four domains: alphabetics, reading fluency, comprehension, and general reading achievement (WWC, 2012a). Within the WWC classification, alphabetics comprises a number of more narrowly defined constructs: phonemic awareness, phonological awareness, letter identification/knowledge, print awareness, and phonics. The WWC comprehension domain comprises vocabulary development and reading comprehension. Because we obtained our effectiveness data from the WWC, for the purposes of our analysis we generally abide by the WWC classification of literacy outcomes but break out the WWC comprehension domain into vocabulary and reading comprehension "domains". We exclude any outcomes that are not strictly reading constructs.

The seven programs we analyze are, like many reading programs, multicomponential, i.e., they each aim to address multiple aspects of literacy and target each literacy domain to varying degrees. Programs for younger children usually place greater emphasis on alphabetics, while programs for older children are more likely to address fluency or reading comprehension. One program may target only alphabetics while others may aim to address multiple literacy domains. It is therefore difficult to compare early literacy programs targeted at different age groups and with different specific goals. Even when the same outcome is targeted across programs, studies of effectiveness often use different measurement instruments. The lack of consistency is problematic: decision-makers are faced with evidence from studies that do not measure effectiveness using the same metrics. Also, to the extent that effectiveness is not measured consistently, cost-effectiveness will not be either. Notwithstanding, we assume that the existing evidence base is the best available for making decisions between programs and interventions to improve literacy.

2.2 Selection Criteria for Early Literacy Interventions to Include in Cost-effectiveness Analysis

For our cost-effectiveness comparison we selected from the WWC inventory of effective early literacy programs using two criteria. First, we identified the subset of the 35 WWC-listed K-3 literacy programs that showed a statistically significant positive impact on at least one test of the phonics construct within the alphabetics domain. Phonics, "the acquisition of letter-sound correspondences and their use in reading and spelling" (NICHD, 2000), was the construct most frequently tested in programs serving K-3 students. Therefore, this selection criterion yields the maximum number of programs with a comparable impact on a literacy construct. The second criterion for inclusion of programs in our analysis was that the WWC accepted a recent evaluation of the program, published since 2005. While this criterion significantly limits the number of programs we study, the purpose of this restriction was to increase the likelihood that we could collect accurate cost data retrospectively. Asking program evaluators, developers or implementers to recall the fine details of program implementation from more than 10 years ago introduces inaccuracies that diminish the value of the cost analysis.

Seven literacy programs met our two criteria: Kindergarten Peer-Assisted Learning Strategies (K-PALS); Stepping Stones to Literacy (Stepping Stones); Sound Partners, Fast ForWord Reading 1 (FFW1); Reading Recovery; Corrective Reading; and Wilson Reading System.[2] Table 1 summarizes key details of each of the

2 One program that met our criteria, Success for All, was excluded because while the seven programs we selected required no significant reorganization of school operations, this program is a whole-school reform model.

Table 1

Program Details for Seven Early Literacy Programs as Evaluated

Program/study characteristic	K-PALS	Stepping Stones	Sound Partners	Fast ForWord Reading 1	Reading Recovery	Corrective Reading	Wilson Reading System
Grade level of students in study	K	K	K	1 and 2	1	3	3
Targeted students in evaluated studies	Average readers	Struggling readers with behavioral disorders	Below average readers, 20-30th percentile	Slightly below average readers	Bottom 20% of readers	Bottom 25th percentile of readers	Bottom 25th percentile of readers
Total number of students receiving intervention	Around 4,400 across 71 schools	65 across 17 schools	54 across 13 schools	103 across 3 schools	94 across 47 schools	44 across 11 schools	53 across 10 schools
Duration	20 weeks	5 weeks	18 weeks	6 weeks	12-20 weeks	28 weeks	28 weeks
Point of impact testing after program start	18 weeks	5 weeks	18 weeks	5-6 weeks	End of intervention and year end	28 weeks	28 weeks
Dosage	35 mins/day, 72 lessons	20 mins/day, 25 lessons	30 mins/day, 4 days/week	43 mins/day, 5 days/week	30 mins/day, 5 days/week	60 mins/day, 5days/week	60 mins/day, 5days/week
Delivery	Whole class with regular teacher, partially replaces regular instruction	1-1 pull-out with tutor, supplements classroom instruction	1-1 or 1-2 pull-out with tutor, supplements classroom instruction	25-30 pull-out students in lab with monitor, supplements classroom instruction	1-1 pull-out with Reading Recovery teacher, supplements classroom instruction	1-3 pull-out with Corrective Reading teacher, supplements classroom instruction	1-3 pull-out with Wilson Reading System teacher, supplements classroom instruction
Evaluation study used for cost-effectiveness analysis	Stein et al., 2008	Nelson, Benner, & Gonzales, 2005; Nelson, Epstein, Stage, & Pierce, 2005	Vadasy & Sanders, 2008	Scientific Learning Corporation, 2005	Schwartz, 2005	Torgesen et al., 2006	Torgesen et al., 2006

Note. In some cases not all of the students receiving the intervention were included in the evaluation study sample due to missing data, sampling procedures or because the students were not randomly assigned to treatment.

seven programs. Within the context of literacy, there are both commonalities and differences. Across the programs, three focused on kindergarten students, two on first-graders, and two on third-graders. One study focused on students of all reading levels, whereas the others focused on struggling or below-average readers. The programs – or more precisely, the versions that were evaluated – operated at different scales, ranging from 44 to 4,400 students, and were each spread across three to 71 schools. Salient for our cost analysis, the programs ranged in terms of: duration (weeks of implementation); dosage (minutes per day); and mode of delivery. These descriptions affirm that, even after restricting our choice set to programs that share similar features, there is considerable variation in literacy programs.

For each of the seven programs that met our criteria, we selected the most recent evaluation study listed by WWC to use as the basis of our cost-effectiveness analysis. It is important for cost-effectiveness analysis to match the costs associated with a specific implementation of a program with the corresponding level of effectiveness observed for that implementation. This generally precludes the use of average effect sizes obtained from multiple evaluations of the same program. In the case of Stepping Stones, two studies that were almost identical in nature and conducted in the same year were combined.[3]

The effectiveness of each program in the domains of alphabetics, fluency, and reading comprehension is reported by WWC in the form of effect sizes known as Hedges' *g*, "the difference between the mean outcome for the intervention group and the mean outcome for the comparison group, divided by the pooled within-group standard deviation on that outcome measure" (WWC, 2013, p.20). Table 2 summarizes the effect sizes provided by WWC intervention reports for each program. It should be noted that effect sizes are more useful for comparison purposes than for direct interpretation, as 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. Table 2, column 3, shows the average annual effect size gain in literacy for different grades as reported by Hill, Bloom, Black, and Lipsey (2007), who argue that effect sizes "should be interpreted with respect to empirical benchmarks that are relevant to the intervention" (p.1). The effect size gains reported for each of the seven programs are based on studies with high quality research designs where a positive effect for at least one literacy construct has been established.[4]

While all the studies from which we obtained effectiveness data employed rigorous research designs involving random or partially random assignment of students to treatment and control conditions,[5] there are several reasons to be cautious in assuming that the effectiveness results observed could be replicated under typical school conditions. In four of seven cases (K-PALS, FFW1, Stepping Stones, and Sound Partners), at least one of the evaluators was also a developer of the program. In the additional case of Reading Recovery, the evaluator has been a trainer for the program and a senior administrator at the Reading Recovery Council of North America. These affiliations with the evaluated programs may introduce bias towards positive

3 Future analyses could include multiple cost-effectiveness assessments for each program, each one based on a different evaluation, in order to obtain a range of cost-effectiveness estimates for each intervention.

4 Six of the studies used random assignment. The evaluation study of Sound Partners uses a quasi-random assignment to guarantee that "each classroom was represented in the control group" and "to ensure a larger dyad-tutored total group size relative to the individual-tutored group size" (Vadasy & Sanders, 2008, p.933).

5 The WWC has established a protocol for evaluating research and it summarizes the evidence from studies that meet reasonable standards of validity, as per its WWC Procedures and Standards Handbook (2013). All of these programs were evaluated by studies that use a randomized controlled trial or quasi-experimental design and meet the requirements of low or moderate attrition, comparable baseline characteristics between the treatment and the control groups in the analytic sample, and appropriate measurement of confounding factors. We expect the resulting estimates of effectiveness to have high internal validity.

results in a variety of ways, not least of which is that fidelity of program implementation is likely to be higher in these situations than in situations where the developer is not actively ensuring the program is delivered as intended.

Table 2
Effect Sizes Observed for Seven Literacy Programs

Program	Grade(s) served	Average annual effect size improvement in literacy for this grade†	Alphabetics	Fluency	Reading comprehension
K-PALS	K	1.52	0.86*	nm	nm
Stepping Stones	K	1.52	0.84na	nm	nm
Sound Partners	K	1.52	0.34ns	0.48*	0.41*
FFW1	1 and 2	0.97/0.60	0.24*	nm	nm
Reading Recovery	1	0.97	0.7*	1.71*	0.14ns
Corrective Reading	3	0.36	0.22na	0.27*	0.17ns
Wilson Reading System	3	0.36	0.33na	0.15ns	0.17ns

Note. Effect sizes from WWC, 2007abcd, 2008, 2010, 2012b; †Hill et al., 2007.

* Statistically significant. na = this effect size is an average of two or more effect sizes at least one of which is statistically significant; ns = not significant; nm = not measured. No results reported for the vocabulary domain because none of the studies we used measured outcomes in this domain.

Furthermore, several of the studies involved significant effort to measure fidelity of implementation (K-PALS, Stepping Stones, Sound Partners, Wilson Reading System, and Corrective Reading), including observations of the instructor working with students and, in some cases, provision of feedback to the instructors to help them improve delivery. Such investments of time by observers are unrealistic in typical school situations such that these studies may not represent the implementation of programs as they are routinely delivered at scale. Compounding this issue is the fact that for all programs but K-PALS, the evaluation involved delivery of the program to a fairly small number of students (in the range of 50 to 100), and it is not clear that similar results could be replicated at greater scale or with different populations.

2.3 Cost Analysis of Early Literacy Interventions

The ingredients method was used to determine the costs of each program (Levin & McEwan, 2001). The purpose behind the ingredients (or resource) approach is to account for the opportunity cost of all of the resources required to implement the particular educational intervention being evaluated, irrespective of their source. By focusing on ingredients, this approach begins not with a budget, but with the details of the intervention and its resource requirements. Budgets are inadequate for accurate cost analysis for several reasons: 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; and they 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 expenditures made (Levin & McEwan, 2001, p.45-46).

The aim of our cost analyses is to estimate the cost of replicating the implementation of each early literacy program in order to achieve impact results similar to those observed in the relevant evaluations. Tsang (1997) emphasizes that a "competent cost analysis can inform decisionmakers about the full resource requirements of an education program, thus avoiding a significant underestimation in costs that can cause difficulties during program implementation" (p.322).

In the evaluation studies we reviewed for the seven literacy programs included in our analysis, most or all of the costs of implementing the program being evaluated were borne by the funding agency sponsoring the study, so that the program was apparently "free" to the schools. We wish to emphasize that the "cost" of a program is determined by the value of the resources that are required, not by how the program is financed. We present costs of replicating program implementation from the perspective of the typical school. We expect that, in typical situations, most of the costs of school-based early literacy programs will be borne by the school itself while some costs, for example, a district-wide literacy coach, might be funded by the school district. Small amounts might be underwritten by families in the form of volunteer time or home-based reading materials. We consider only the costs of the programs above and beyond the resources students already receive as part of their regular instruction in school, i.e., we identify the incremental costs of introducing the programs into existing school activities.

Each program we studied displaced some other instruction for the students receiving the intervention. In most cases where a few students were pulled out of the main classroom to participate in a supplementary literacy program, we determined that there were unlikely to be any significant changes in instruction in the main classroom from which they were temporarily removed. The slightly reduced class size would still likely have required the same number of personnel and use of facilities. It is possible that slightly fewer materials were utilized in the main classroom but as these are generally a tiny percentage of costs, they would not significantly impact overall costs.

One program, K-PALS, was delivered to the whole classroom by the regular classroom teacher, in the same classroom space, as a partial substitute for regular reading instruction. We assumed that this substitution neither added to nor subtracted from the costs of the teacher and facilities for instructional time. However, if we were able to determine the precise ingredients used during regular reading instruction and their costs, we would be able to assess whether K-PALS actually cost more or less than the programming it replaced. Again, most likely the differences would be in materials and equipment which account for a small proportion of most of the interventions we review. We are also not able to account for the costs of lost regular instructional time because assessments of outcomes beyond literacy were not included in the evaluations. For example, if students were regularly pulled out of science classes to participate in a reading intervention, they would probably perform less well on assessments of science achievement.

An initial list of the ingredients required to implement each program was compiled through careful review of evaluation studies listed by WWC and other publicly available articles, reports, web sites or materials for each program. A detailed interview protocol was developed for each program (based on a generic protocol we devised, see Appendix II) to elicit further information regarding the ingredients identified and to identify additional ingredients not already listed. 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 people involved in implementing the program, whether directly or peripherally. For example, while an evaluation report may have indicated that tutors were used to deliver a program

four times a week in one hour sessions, we collected detailed information about the qualifications and work experience of the tutors, what proportion of their work time was spent on the program, and how many hours were spent in training, preparing lessons, tracking student progress, communicating with the classroom teacher, principal, parents, and so on.

We contacted the developers and the evaluators of each program, inviting them to participate in telephone interviews to answer questions about the program ingredients. Depending on the complexity of the program and the resource details already available prior to the interviews, the interviews ranged in length from 40 minutes to 2 ½ hours. Follow up questions or clarifications were answered through brief phone calls or via email. In each case we also asked whether we could obtain identities of the schools and teachers or trainers who participated in the evaluations so that we could obtain site-level ingredients data and investigate how implementations may have varied across sites. In most cases, the evaluators' confidentiality agreements with study participants precluded this possibility. However, we were able to interview one or more persons beyond the evaluators who were (or are) directly involved in implementations of FFW1, Corrective Reading, Wilson Reading System, and Reading Recovery.

Once the ingredients required to implement each program were specified, the next step was to associate each ingredient with a national price to make the programs directly comparable. Most prices were obtained from publicly available databases such as the National Occupational Employment and Wage Estimates by the Bureau of Labor Statistics. Appendix III provides details on our sources for national prices. In some instances, we used a specific price obtained from the program developer such as the cost of a site license for FFW1. All prices are converted to 2010 dollars for consistency across programs although a few materials and equipment items such as computers are in current dollars as price changes do not occur in line with inflation and/or 2010 prices are not easily available. All costs associated with initial training to implement a program are amortized over 3 years except for situations where we know the average tenure of the personnel receiving the training, in which case we amortize over the period of tenure. We do not amortize ongoing professional development that occurs on a regular basis. For educational facilities rental rates are not generally available as national prices, so we use total construction costs of school buildings (construction costs adjusted for cost of land, development, furnishings and equipment) and amortize over 30 years. We use a 3% interest rate for amortization, reflecting the current yield of 30 year U.S. Treasury Bonds. Using a higher interest rate (e.g., 5%) yields higher per student costs for facilities, but because in all cases facilities costs are a small percentage (up to 7%) of the total, the relative costs of the programs are not highly sensitive to the interest rate used.

Costs for all of the programs except Reading Recovery reflect the program as evaluated in the studies we selected. Costs for Reading Recovery are based on an "average" implementation as described by the developers and evaluators of the program because we were not able to identify an interviewee who could recall enough details about the evaluated implementation and insufficient information was available in written reports.

For five of the programs (K-PALS, Stepping Stones, Sound Partners, Corrective Reading, and Wilson Reading System), significant resources were devoted towards assuring fidelity of implementation such as having trained observers watch lessons being delivered and providing feedback to the instructors. Any activities that we believe may have affected the impact of the program were included as a cost while those that were associated only with the research requirements of conducting an evaluation were not included. For example, administration of post-tests was not counted as a program cost if the purpose was simply to determine program impact. However, if the post-tests were used to determine continuation in the program we did include the associated costs. Pre-tests were counted as a cost if they were used as screening measures to determine treatment eligibility or placement.

2.4 Comparability of Early Literacy Interventions

Differences in literacy outcomes targeted

Even after applying our selection criteria to facilitate a cost-effectiveness comparison of early literacy programs, we still faced a number of methodological and empirical challenges with respect to their comparability. First, the seven programs were each designed to improve a variety of early literacy domains and constructs, not only phonics. In fact, according to the developer of Stepping Stones, the program does not target phonics skills directly but places an emphasis on phonological and phonemic awareness, important precursor skills shown to have causal impacts on phonics (NICHD, 2000). In some of the evaluations, measures were used to assess impact on a literacy construct that the program did not aim to address, and in some cases the evaluation did not assess impact on all the constructs that were addressed. These inconsistencies and gaps in measurement of effects are problematic when attempting to compare programs for overall impact on literacy.

Differences in the number of literacy outcomes addressed by a program should be considered when evaluating efficiency of resource use because in some cases the investment is "buying" more than one outcome. To address this issue we collected data from program developers and evaluators on the average percentage of program delivery time that was allocated to each literacy construct/domain, summarized in Table 3. We subsequently distribute costs for each program across the literacy domains targeted by the program using the proportions from Table 3. To facilitate comparability of outcomes among the programs, we aggregated the more granular literacy constructs on the survey into the four overarching domains of alphabetics, passage reading fluency (hereafter referred to as fluency)[6], vocabulary, and reading comprehension.

The evaluator of Reading Recovery did not feel that the program goals could be parsed into individual constructs or domains because "the various criterion measures are very interrelated and just provide an indication of developing processing systems for reading and writing" (R. M. Schwartz, personal communication, February 19, 2013). While the program addresses all components of early literacy, the emphasis on each varies according to each individual student's needs (Schwartz, 2005). WWC reports impact findings for Reading Recovery in the alphabetics, fluency, and reading comprehension domains.[7] We assume, for the purposes of our cost-effectiveness calculations, that Reading Recovery targets each of the four domains of alphabetics, fluency, vocabulary, and reading comprehension equally in order to distribute program costs across the multiple outcomes. We assume that the emphasis on each of these domains will vary by individual student, but that they receive roughly equal amounts of emphasis when the instructional efforts are aggregated across children. We recognize the limitations of this assumption in that it may not perfectly capture the integration of elements in Reading Recovery. However, we concluded that it was the most reasonable approach to allow us to incorporate the information on Reading Recovery in our study. We also provide an alternate analysis in which only 10% of delivery time is attributed to the alphabetics domain to demonstrate the impact of changing this assumption on the cost-effectiveness ratio. Future research should investigate alternate approaches to parse the instructional emphasis on different domains for multicomponential literacy programs.

6 It is worth noting that we followed the WWC classification scheme in classifying the construct of word reading efficiency under the domain of alphabetics, and distinguishing this from the domain of passage reading fluency. We recognize that other researchers may group word reading efficiency and passage reading fluency together and encourage future research to consider this possibility, but we believe this is a reasonable decision in alignment with the WWC classification of outcome measures.

7 In addition, findings are reported for general reading achievement but, as the measurement instruments indicate the actual outcomes measured are writing concepts and not reading constructs, we do not include them in our analysis.

Table 3

Percent of Program Delivery Time Addressing Literacy Outcomes

Literacy domain	Literacy construct	K-PALS	Stepping Stones	Sound Partners	FFW1	Reading Recovery	Corrective Reading**	Wilson Reading System
Alphabetics	Phonological/Phonemic Awareness	20%	55%	5%	15%	—	2.5%	12%
	Phonics/Decoding	20%	—	50%	10%	—	55.5%	21%
	Sight Word Recognition	20%	—	15%	10%	—	—	9%
	Word Reading Fluency	15%	—	10%	10%	—	25%	14%
	Letter Identification/Knowledge	—	30%	10%	5%	—	—	5%
	Print Awareness/Print Concepts/Print Knowledge	—	10%	—	—	—	—	5%
	Other (Letter Sound Correspondence)	10%	—	—	—	—	—	—
	Other (Rapid Naming)	—	5%	—	—	—	—	—
Total for Alphabetics		85%	100%	90%	50%	25%	83%	66%
Fluency	Passage Reading Fluency	15%	—	10%	10%	25%	17%	<2%
Vocabulary	Vocabulary	—	—	—	18%	25%	—	—
Reading Comprehension	Reading comprehension strategies	—	—	—	15%	25%	—	4%
*	Other (Spelling/Encoding)	—	—	—	5%	—	—	28%
*	Other (Grammatical concepts)	—	—	—	2%	—	—	— —
Total		**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Note. Source of data was survey responses from program developers, except for Reading Recovery where we assign an equal percentage of time to each of the four literacy domains. Wilson Reading System and Corrective Reading were modified for the evaluated implementation to eliminate the comprehension strand for Corrective Reading and the vocabulary and comprehension components for Wilson Reading System. Typical implementations of these programs include more time devoted to these literacy domains. * We do not show a domain for these items because they are considered to be writing concepts rather than reading concepts. **For Corrective Reading the developer gave us percent time for each of the two program levels (B1 and B2) used by the majority of students in the Torgesen et al. (2006) study. We averaged the data for the two levels assuming that an equal amount of time was spent on each.

Program Differences in Duration, Target Population, and Assessment Measures

In addition to considering the variety of literacy outcomes targeted among programs, a second concern relates to comparability across programs with different durations. While individual students in all program evaluations we relied on for effectiveness data benefited from the interventions for less than one year, treatment duration varied substantially. FFW1 and Stepping Stones were delivered over just 5-6 weeks, Reading Recovery lasted 12-20 weeks per cohort, Sound Partners lasted 18 weeks, K-PALS lasted 18-20 weeks per cohort, and both Corrective Reading and Wilson Reading System were implemented over 28 weeks. Outcomes were generally measured at the end of the intervention period, except for Reading Recovery where outcomes were measured again six months after the intervention ended for the first cohort treated. We note that a number of empirical studies either raise questions or provide evidence with respect to fade-out of effects over time, e.g., Rothstein (2010), for teacher effectiveness; Camilli, Vargas, Ryan, and Barnett (2010), and Currie (2001), for early childhood interventions. Fade-out effects may vary based on student characteristics and program design features (Currie & Thomas, 1995). None of the seven literacy programs we reviewed conduct longer term follow-up testing to investigate the fade-out effects.

A third comparability issue relates to grade spans and student abilities. While each of the seven programs was implemented with students in grades K-3, the target populations differed within this range, and generalizability of the results to other populations is uncertain. The K-PALS study targeted all students in regular kindergarten classrooms, including struggling, average, and high-achieving readers. The Stepping Stones study also targeted kindergarten students, but only those who were at risk for behavioral disorders and reading difficulties. Sound Partners targeted kindergarten students scoring between the 20th and 30th percentiles of standardized reading tests. FFW1 targeted first and second graders who were slightly below average readers, and Reading Recovery targeted the bottom 20% of readers of each first grade classroom. Corrective Reading and Wilson Reading System were both used with third grade students, scoring mostly in the bottom 25th percentile of their grade on standardized tests.

Even across these grade spans, it is debatable whether literacy gains are comparable. Hill et al. (2007) report that annual gains in reading test scores vary substantially but predictably across grades: the average reading test effect size for students progressing through kindergarten to first grade is 1.52 compared with 0.97 for Grade 1-2 students, and 0.36 for third-fourth graders. Programs targeting older students may look less effective than those targeting younger students, even if the students show a greater than typical gain in literacy. This suggests that a fair comparison of reading programs can only be made by including studies of programs addressing students in the same grade. Alternatively, if the data were available, cross-grade-level comparisons might be viable if observed gains could be compared to the expected gains for each given grade level and reading level of participants (e.g., struggling or average), in order to determine the "value-added" gain provided by each program.

A fourth concern is that the measures used to capture program impacts were not identical across studies, even for the same domain, and it is possible that some measures are more sensitive to instruction than others. For example, the studies of K-PALS and FFW1 used Rapid Letter Sounds and TOPA Letter Sounds respectively, both of which are measures that are arguably more proximal to the phonics instruction provided, while the studies of the other five programs used more distal measures. Use of proximal measures may increase the likelihood of a large effect size. Hill et al. (2007) report that while the average effect size from 389 random assignment studies of interventions targeting elementary school students is 0.33, the range extends as low as 0.07 when the measures of effectiveness are "broadly focused" standardized

tests (p.8), (analogous to a distal measure in our terms), to as high as 0.44 for specialized tests (proximal measures in our terms).

Effect sizes are also likely to be inflated when the study participants are relatively homogenous with respect to performance on the measure of outcome, in this case when reading abilities are similar (Fern & Monroe, 1996; Olejnik & Algina, 2000). If the standard deviations for the outcome measure scores are small (because students are more homogenous), the effect size will automatically appear larger than for a group of participants who score more diversely. A study focused only on struggling readers scoring below a certain percentile on a reading test is therefore prone to effect size inflation relative to a study of whole classrooms.

3. COST-EFFECTIVENESS ANALYSIS OF SEVEN EARLY LITERACY INTERVENTIONS

3.1 Overview and Limitations of Cost-effectiveness Analysis

The primary purpose of cost-effectiveness analysis is to compare interventions in terms of efficiency of resource use in addressing a common outcome. 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 this study we divide cost per student, adjusted for percent of program delivery time addressing each literacy domain, by the effect size (reported in Table 2) observed on measures of alphabetics, fluency, or reading comprehension. 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.

Unfortunately, because site-level sample sizes are so small in almost all the program evaluations we use for the seven programs, it is not possible to investigate site-specific variation in cost-effectiveness, except in the case of FFW1. Levin et al. (2012), and Bowden (2013) demonstrate that cost-effectiveness ratios can vary widely for the same program being implemented at different sites. These differences may be a result of variation in use of ingredients, the efficiency of their usage, prices for a given ingredient, and to differential effectiveness in delivering the program to the specific population being served at each site. We were able to obtain both site-level effectiveness and site-level cost data only for FFW1. For K-PALS and Sound Partners we were able to obtain costs and effects for variations in the implementation, but not at the site level. For Reading Recovery we were able to obtain cost data for two different sites but not site-level impact data. For all programs, our interviewees were able to give us some indications of how implementation might vary across sites and we used this information to conduct sensitivity analyses to investigate how these variations might impact costs.

A further issue for our analysis created by data limitations relates to the issue of incremental cost-effectiveness. In all the studies we reviewed, students received the intervention during school time so that time was lost on other academic activities. In some cases the intervention partially replaced existing literacy instruction, while in others, students were pulled out of periods devoted to social sciences or other subjects. Ideally, evaluators should attempt to assess whether other outcomes were affected even as some literacy outcomes may have been improved. Levin (2002) analyzes comprehensive school reforms and observes that shifting resources from an unmeasured outcome to a measured outcome will likely improve results on the measured outcome, but a full accounting of the costs and effects should take into account any offsetting losses on unmeasured outcomes. In the case of the literacy programs we study, the evaluators could administer pre- and post-tests for other literacy outcomes and subject areas to determine whether the students receiving the literacy intervention performed any differently from the comparison group receiving "business as usual". Because these assessments were not made, we assumed that the literacy effects reported for each program were additional to effects from business as usual.

While our initial intention was to compare all seven programs in a cost-effectiveness analysis, the differences noted above in age group and reading level served, outcomes measured, and duration of interventions suggest that more limited comparisons are prudent. For each program we first perform an individual analysis: we describe the intervention, report on its resource usage and cost of delivery, and then relate these costs to the evidence on effectiveness to derive cost-effectiveness ratios. We also perform sensitivity testing. Subsequently, we offer direct comparisons only among programs targeting same-grade students of comparable reading ability.

K-PALS, Stepping Stones and Sound Partners all target kindergarten students, but K-PALS stands out because it is delivered to entire classrooms of mixed ability readers as opposed to being a pull-out, small group program for struggling readers. Hence we compared the two programs for struggling readers to each other (Stepping Stones and Sound Partners) and suggest that K-PALS is not a directly comparable alternative to these two programs. Even the comparison of Stepping Stones and Sound Partners has limitations because the former lasts 5 weeks and the latter 18 weeks.

Reading Recovery and FFW1 both target first grade students who are below-average readers, although the study of FFW1 also included some second graders. Cost-effectiveness ratios of these two programs may also be compared although FFW1 is much shorter in duration (6 weeks vs. 12-20 weeks for Reading Recovery).

Both Wilson Reading System and Corrective Reading were studied in the same evaluation. The programs were implemented with third grade struggling readers for the same length of time, and using the exact same outcome measures to test for impact on literacy constructs. This situation provides the best-case scenario for application of cost-effectiveness analysis.

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: 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. For each program we analyze, we consider one or more sensitivity tests to evaluate the impact of alternate assumptions on the cost-effectiveness ratios we present. Because the programs are all short in duration (less than one year), there is no need to adjust costs for their time value, therefore we do not report sensitivity tests that vary the discount rate. As previously mentioned, varying the interest rate used for amortization also has little impact on costs because items that are amortized over long periods of time, such as facilities, account for only a small percentage of the program costs for the interventions we studied. The major variables and assumptions we test and report are: mode of implementation, number of students served by each program or by each instructor, and the costs of the personnel who represent the most significant expense.

3.2 Kindergarten Peer-Assisted Learning Strategies

K-PALS Program Description and Effectiveness

K-PALS is one of several iterations of Peer-Assisted Learning Strategies created by L. and D. Fuchs at the Vanderbilt Kennedy Center for Research on Human Development at Vanderbilt University. The program is used in mainstream kindergarten classrooms by the entire class. At the beginning of each session, the teacher models being a coach to the entire class, guiding the group through a phonics game or exercise. The class subsequently breaks into pairs, with each student in the pair taking turns being the "reader" and "coach." The coach guides the reader through exercises while giving feedback and encouragement. K-PALS falls under the umbrella of a collection of peer-assisted learning programs (PALS) developed by the same researchers for improving reading and mathematics outcomes; the other PALS programs incorporate similar instructional routines (e.g., pairing) but with different, age-appropriate curriculum for different grade levels of students. Other implementations of PALS, such as a program targeting reading achievement for English language learners in third through fifth grades, have been identified as having "potentially positive" effects by the WWC.

The evaluation study that we selected for our analysis was conducted by Stein et al. (2008) from 2004-2005. The program developers were among the evaluators conducting the study. The study sample included 3,171 students from 71 schools in Minnesota, South Texas, and Nashville, Tennessee. Prior to program delivery, all students in the treatment classrooms were screened using the Rapid Letter Sounds test, assigned within classrooms to "High," "Medium" and "Low" reading ability and ranked in order of score. Classroom teachers used the ranked scores to pair students so as to create a consistent and optimal difference in achievement between students within pairs. The classroom teachers received one day of training in how to use K-PALS in their classrooms and subsequently embarked on a series of 72 lessons, 35 minutes each, delivered over 18-20 weeks. Student pairs were changed every 4-6 weeks.

In addition to the base-case "workshop" implementation with only one day of training, two additional levels of implementation provided increasing levels of training and support for the classroom teachers in their delivery of K-PALS. At the second level, the classroom teachers received two additional training sessions, or "boosters" during the treatment period. At the third level, the teachers were provided with the additional booster training sessions and also with graduate student "helpers" in the classroom. The evaluators wanted to investigate how additional training and support might affect efficacy of the program. Throughout the two-year study, fidelity of implementation measures were employed to assess whether teachers were delivering the program as intended.

After 17 weeks of lessons, the program impact on students' phonics skills was measured using the Rapid Letter Sounds test. As noted previously, this test may be particularly sensitive to instruction such that the large effect sizes observed for each implementation level might not be replicable if more distal measures of phonics were used. The base-case "workshop" implementation with only one day of training for the classroom teachers resulted in an effect size of 0.61 for alphabetics skills. The booster and helper levels of implementation yielded effect sizes of 1.07 and 0.89 respectively for alphabetics skills. It appears that while the booster sessions helped improve the impact of the program, the addition of a helper in the classroom did not add further to the impact, but in fact reduced it from the booster level.

K-PALS Ingredients and Costs

Through phone interviews and email exchanges with a K-PALS developer and evaluators at Vanderbilt University we created a list of the ingredients utilized in program delivery. While all three implementation levels utilized the same baseline ingredients, the booster level required additional training time and the helper level further added a graduate student. Table 4 lists the ingredients at each level of implementation, their estimated total costs for the period of the evaluation, the cost per student of each ingredient, and the percentage of total costs accounted for by each of the categories of ingredients (personnel, facilities, materials and equipment, and other inputs).

The program was used with entire classrooms but only the four top, four bottom and four middle scorers on the initial screening test were included in the study sample. The total number of students served by the program was not available so, in order to spread costs across all students in the K-PALS classrooms, we estimated the total number of students as follows: we used the reported proportion of sample students at each implementation level and the total number of teachers reported to estimate the number of teachers in each condition (control, workshop, booster, helper); we then assumed that each teacher taught a class of 20 students, using the midpoint of the class size range of 16-24 students reported to us by one of the program evaluators.

For the workshop level of implementation, total costs for 1,732 students across 71 schools were $46,566 and costs per student were very low at $27, with 60% being attributable to personnel. It is notable that

Table 4
Ingredients and Costs for K-PALS

	Workshop level			Booster level			Helper level		
Ingredient	Ingredient cost	Cost per student	Percent of total	Ingredient cost	Cost per student	Percent of total	Ingredient cost	Cost per student	Percent of total
Personnel total	**$27,938**	**$16**	60%	**$41,646**	**$25**	68%	**$53,796**	**$51**	79%
Classroom teacher	$9,648	$6		$15,156	$9		$9,796	$9	
Substitute teachers	$3,966	$2		$3,738	$2		$2,416	$2	
Screening testers	$1,440	$1		$1,374	$1		$845	$1	
Project Coordinators – training	$2,332	$1		$2,224	$1		$1,429	$1	
Project Coordinators – implementation fidelity	$10,552	$6		$10,065	$6		$6,466	$6	
Principal Investigator	—	—		$7,131	**$4**		$4,617	$4	
Graduate assistants (providing training)	—	—		$1,958	$1		$1,268	$1	
Project staff – training and supervising helpers	—	—		—	—		$5,939	$6	
Graduate student helpers	—	—		—	—		$21,020	$20	
Facilities total	**$497**	**$0**	1%	**$1,410**	**$1**	2%	**$2,283**	**$2**	3%
Classroom for training	$497	$0		$474	$0		$305	$0	
Classroom for booster	—	—		$935	$1		$618	$1	
Classrooms for helper training and meetings	—	—		—	—		$1,361	$1	

(continued on next page)

Table 4
Ingredients and Costs for K-PALS (*continued*)

	Workshop level			Booster level			Helper level		
Ingredient	Ingredient cost	Cost per student	Percent of total	Ingredient cost	Cost per student	Percent of total	Ingredient cost	Cost per student	Percent of total
Materials and equipment total	**$10,301**	**$6**	22%	**$9,781**	**$6**	16%	**$6,297**	**$6**	9%
Copies – Rapid Letter Sounds	$67	$0		$64	$0		$41	$0	
K-PALS manual	$1,190	$1		$1,122	$1		$725	$1	
Large print lessons	$1,263	$1		$1,190	$1		$769	$1	
DVD	$17	$0		$16	$0		$10	$0	
Student score sheets	$4,826	$3		$4,603	$3		$2,956	$3	
Overhead projector	$564	$0		$531	$0		$343	$0	
Transparencies, pens, folders, stickers	$2,375	$1		$2,254	$1		$1,451	$1	
Other inputs total	**$7,830**	**$5**	17%	**$8,647**	**$5**	14%	**$5,539**	**$5**	8%
Coordinator travel – implementation fidelity	$6,583	$4		$6,279	$4		$4,035	$4	
Coordinator travel – training	$1,247	$1		$1,189	$1		$741	$1	
Local travel to booster sessions	—	—		$1,178	$1		$762	$1	
Grand total	**$46,566**	**$27**	100%	**$61,483**	**$37**	100%	**$67,915**	**$64**	100%

almost $10 of this cost per student was accounted for by the fidelity of implementation checks which may not occur at all or be as rigorous in a typical implementation of K-PALS, such that the total cost per student could drop to as low as $17 per student. However, it is possible that without the ongoing supervision and feedback, teachers would not implement the program as faithfully as they did in this study, resulting in lower impact on students' literacy skills than observed here.

K-PALS Cost-Effectiveness and Sensitivity Analysis

Table 5 shows how costs, effects and cost-effectiveness varied at the three different implementation levels. The only literacy construct measured was phonics, which is in the alphabetics domain. The program developer indicated that 85% of K-PALS delivery time targeted constructs in the alphabetics domain, while 15% targeted the fluency domain. We therefore apply 85% of the costs of the program to the effect size observed for alphabetics at each level of implementation. The cost-effectiveness ratios represent the incremental cost to obtain one standard deviation increase in alphabetics skills. Across all three implementation levels, K-PALS effect size gains are obtained at a very low cost. The addition of the booster training sessions appears to be a productive use of resources as the cost-effectiveness ratio improves from $38 to $29 per standard deviation increase in alphabetics skills. However, the further addition of a graduate student helper in the classroom to assist the teachers does not appear to be a worthwhile use of resources as the cost-effectiveness ratio rises to $61 per standard deviation increase in alphabetics skills.

As explained above, we made an assumption that average class size served by K-PALS was 20 students. If average class size were lower than this, the K-PALS costs per student would be higher as few of the costs vary by class size, while if class size were higher, the bulk of the costs would be spread out over the larger population served. Varying only the variable costs by the number of students served in the workshop level K-PALS program, the cost per student would increase from $27 to $32 if average class size were 16 students and would drop to $23 if average class size were 24 students. If we could assume that the program impact is unchanged with class size, the cost-effectiveness ratio moves only modestly from $38 to $45 per standard deviation increase in alphabetics skills for a class size of 16 and to $32 for a class size of 24. If longer term effects were measured for the program, for example, a year after program exposure ended, it would be valuable to determine whether any effect size changes might alter the ranking of the cost-effectiveness ratios for the three different implementation levels of K-PALS.

Table 5
Cost-effectiveness of K-PALS for Alphabetics

Implementation level	Effect size for alphabetics	Cost per student attributed to alphabetics	Cost per unit increase in effect size
Workshop level	0.61	$23	$38
Booster level	1.07	$31	$29
Helper level	0.89	$54	$61
Sensitivity analysis:			
i) Workshop level – smaller class size	0.61*	$27	$45
ii) Workshop level – larger class size	0.61*	$20	$32

Note. Effects from WWC 2012b. 85% of program costs from Table 4 are attributed to the alphabetics domain based on the program developer's indication of percent of program delivery time addressing constructs in this domain, see Table 3. * These effect sizes are assumed to apply for the sensitivity analyses.

3.3 Stepping Stones to Literacy

Stepping Stones Program Description and Effectiveness

Stepping Stones is a tutoring program that was developed in 2004 by R. Nelson, P. Cooper, and J. Gonzalez at the University of Nebraska-Lincoln. The program is focused on phonemic awareness and is "soft-scripted." It is designed to require minimal training and planning to implement successfully. It consists of 25 guided lessons led by a tutor, paraprofessional, or volunteer on a one-to-one basis with underachieving readers in kindergarten and above. Students are pulled out of class to work with the tutor. Lessons include brief instructional activities on such pre-reading skills as identifying environmental sounds, naming letters, and recognizing rhymes. We assume that all costs of Stepping Stones are incremental beyond the regular costs of schooling. Implementation of Stepping Stones in schools can vary along several dimensions; for example, in some cases, teachers deliver the program to small groups of students, instead of tutors administering the program one-on-one.

WWC lists two recent evaluations of Stepping Stones: Nelson, Benner, and Gonzalez (2005), and Nelson, Stage, Epstein, and Pierce (2005). These studies show positive outcomes for three alphabetics constructs: phonological awareness, phonics, and letter knowledge. At least one of the evaluators in each study was also a program developer. The studies were almost identical with one involving 10 schools in a Midwestern city and the other involving 7 moderate- to high-poverty schools in a medium-sized Midwestern city. Given their similarity in almost all respects, we combine the studies and treat them as one. Both implementations targeted kindergarten children who were at risk for behavioral disorders and reading difficulties, with a total of 65 students treated.

WWC reports an average effect size for Stepping Stones in the alphabetics domain, computing it as a simple average effect size for the 65 students across seven measures and two studies: 0.84. Effect size is reported for the overall group because the small number of students at each site (less than four on average) precluded viable site-level analysis.

Stepping Stones Ingredients and Costs

We obtained information on the ingredients used in the evaluated program implementations of Stepping Stones from an individual who was both a program developer and evaluator. Site identities of the 17 schools were not available so we were not able to interview actual program implementers (teachers/tutors) to determine whether and how ingredients varied across schools. Table 6 lists the Stepping Stones ingredients and their associated costs. The cost per student of the program was $479, of which 95% was associated with personnel, mostly for the tutor delivering the program. Facilities costs for the program are quite low as it is short in duration and requires only enough space for the tutor and a single student. The developer indicated that the space requirements for the program are quite flexible and it can be delivered in the back of a classroom, a corner of the school library, or an empty office. Our estimate of program costs is likely conservative, as we assumed that tutors implementing Stepping Stones in the evaluation setting would spend the full school day on Stepping Stones; if the program were replicated, the costs might be lower if tutor time were more flexible. Further, some of the costs, such as for observers, might not be incurred in a regular program implementation. However, the effectiveness of the program might be lower without these observations and the consistency afforded by having a handful of tutors devoted to the program.

Table 6
Ingredients and Costs for Stepping Stones

Ingredient	Total cost (65 students)	Cost per student	% of total
Personnel total	**$29,360**	**$452**	**95%**
Tutor	$26,616	$409	
Tutor – Training	$235	$4	
Tutor Trainers	$517	$8	
Classroom Teacher	$861	$13	
Observers	$1,131	$17	
Facilities total	**$244**	**$3**	**1%**
Classroom for training	$34	$0	
Space for tutoring	$210	$3	
Materials/equipment total	**$1,535**	**$24**	**5%**
Lesson Book	$1,474	$23	
Photocopies for screening	$60	$1	
Grand total	**$31,138**	**$479**	**100%**

Stepping Stones Cost-Effectiveness and Sensitivity Analysis

Based on the program developer's indication (see Table 3) that 100% of Stepping Stones delivery time addresses the alphabetics domain, we attribute all of the costs of the program to the 0.84 effect size gain. The incremental cost-effectiveness ratio, i.e., the cost per standard deviation increase in alphabetics skills, is $570, as shown in Table 7.

The Stepping Stones program developer indicated that some schools use teachers rather than tutors to deliver Stepping Stones and that in some cases the teachers work with small groups rather than one-on-one. We modeled a variation in the program implementation in which the 34 classroom teachers, who were only peripherally involved in the evaluated implementations, taught four students each themselves, instead of tutors teaching students individually. The cost per student drops from$479 to $151. Although teacher time is more expensive than tutor time, we assume that teacher time is used more efficiently throughout the day given their existing teaching commitments, and therefore allocate less transition time to teachers than we do for tutors. Additionally, more students are being served at once. If the effect size of 0.84 for alphabetics skills still holds in this altered scenario, the incremental cost-effectiveness ratio for alphabetics would drop sharply from $570 to $180 per standard deviation increase in alphabetics skills.

We assumed that costs are essentially the same for both implementations of Stepping Stones (Nelson, Benner, et al., 2005; Nelson, Stage, et al., 2005) given the similarity of the studies, but, as an additional sensitivity analysis, we calculate the cost-effectiveness ratio for alphabetics using the two separate effect sizes for each study. This demonstrates the variation in effectiveness and cost-effectiveness masked by averaging effect sizes across studies. Note that, in most situations, costs could not be assumed to be the same across implementations so that both costs and effects would generally be expected to vary between studies of the same program. For the Nelson, Benner, and Gonzalez (2005) study, the effect size was 0.8, resulting in a cost-effectiveness ratio of $599 per standard deviation increase in alphabetics skills. For the Nelson, Stage,

Epstein, and Pierce (2005) study, the effect size was 0.87, yielding a cost-effectiveness ratio of $551 per standard deviation increase in alphabetics skills.

Table 7
Cost-effectiveness of Stepping Stones for Alphabetics

Program	Effect size for alphabetics	Cost per student attributed to alphabetics	Cost per unit increase in effect size
Stepping Stones	0.84	$479	$570
Sensitivity analysis:			
i) Teachers as instructors	0.84*	$151	$180
ii) Nelson, Benner, & Gonzales study only	0.80	$479	$599
iii) Nelson, Stage, Esptein, & Pierce study only	0.87	$479	$551

Note. Effect sizes from Table 2. Costs for Stepping Stones are from Table 6. 100% of program costs are attributed to the alphabetics domain based on the program developer's indication of percent of program delivery time addressing constructs in this domain, see Table 3. *This effect size is assumed to apply for this sensitivity analysis.

3.4 Sound Partners

Sound Partners Program Description and Effectiveness

Sound Partners, developed by the Washington Research Institute in 2004, is a phonics-based tutoring program that provides supplemental reading instruction to K–3 students with below-average reading skills. The program is designed specifically for use by tutors with minimal training and experience. The program consists of a set of scripted lessons in alphabetics and phonics skills and uses BobBooks® beginning reading series as one of the primary texts for oral reading practice. The tutoring can be provided as a pull-out or after-school program, as well as by parents who homeschool their children.

While the WWC lists several evaluations of Sound Partners, we base our analysis on the most recent study by Vadasy and Sanders (2008) from the Washington Research Institute (the developer of the program). The program was implemented across 30 classrooms in 13 urban public schools with 54 kindergarten students considered at-risk for reading difficulties. One-to-one or one-to-two (dyad) pull-out sessions were conducted with the students for 30 minutes a day, four days per week over a period of 18 weeks.

Vadasy and Sanders (2008) employed multiple measures to assess the impact of Sound Partners on various literacy constructs. In the alphabetics domain, the impact was mixed with a statistically significant positive effect size for phonological awareness (0.59) and for one measure of phonics, the WRMT-R Word Reading Accuracy test (0.63). However, a non-significant effect size for letter knowledge (-0.14) and for another measure of phonics, Composite TOWRE: Phonemic Decoding Efficiency and Sight Word Efficiency (0.29), resulted in an overall effect size for the alphabetics domain of 0.34, which according to WWC is not significant. WWC also reports a significant positive effect size for the fluency domain (0.48) and for the reading comprehension domain (0.41).Two other recent studies of Sound Partners listed by WWC report varying effects sizes for these same domains. Vadasy, Sanders, and Peyton (2006), report an effect size of

0.45(ns) for alphabetics, 0.8 for fluency, and 0.28 (ns) for reading comprehension. Vadasy, Sanders, and Peyton (2005), report an effect size of 0.85 for alphabetics, 0.67 (ns) for fluency, and 0.75 (ns) for reading comprehension. While we only conduct a cost-effectiveness analysis using the 2008 study, future work could include similar analyses using the 2005 and the 2006 studies to investigate the range of resulting cost-effectiveness ratios for the various literacy domains.

Sound Partners Ingredients and Costs

We obtained information on the ingredients utilized in the implementation of Sound Partners from the evaluation study, from e-mail communications with the evaluator, who was also one of the developers, and from a telephone interview with the same individual. As shown in Table 8, 72% of the total $791 costs per student are accounted for by the tutors delivering the program. In addition, over $90 per student are attributed to the costs of ensuring fidelity of implementation. These fidelity-of-implementation costs would most likely be significantly lower or absent in a typical implementation of the program, although without these measures, the effect size might also be lower.

Table 8
Ingredients and Costs for Sound Partners

Ingredients	Total Cost (54 students)	Cost per student	% of Total
Personnel total	**$38,400**	**$711**	**90%**
Developer	$1,252	$23	
Project coordinator	$1,700	$31	
Observer	$2,283	$42	
Principal	$2,211	$41	
Teachers for student selection	$554	$10	
Tutor	$30,399	$563	
Facilities total	**$581**	**$11**	**1%**
Tutoring space	$550	$10	
Training space	$31	$1	
Materials and equipment total	**$2,925**	**$54**	**7%**
Screening materials	$18	$0	
Sound Partners Package	$2,675	$50	
Supplies	$194	$4	
Training equipment and materials	$37	$1	
Other inputs total	**$798**	**$15**	**2%**
Transportation	$798	$15	
Grand Total	**$42,703**	**$791**	**100%**

Based on the program developer's indication that 90% of program delivery time addressed alphabetics (Table 3), we attribute 90% of program costs to this domain. Using the average effect size of 0.34 reported by WWC, the cost-effectiveness ratio for the alphabetics domain, as shown in Table 9 (Tutoring vs. Control band) is $2,093 per standard deviation increase in alphabetics skills. While we would not generally calculate a cost-effectiveness ratio for an outcome showing a non-significant effect size, we did so here because the WWC-reported effect size is an average of several effect sizes, two of which are statistically significant. For the fluency domain, the cost-effectiveness ratio is $165 per standard deviation increase in fluency skills. This is lower than the cost-effectiveness ratio for alphabetics because we attribute only 10% of the program costs to the fluency domain (i.e., $79) and the effect size (0.48) is larger than for alphabetics.

WWC also reports a positive impact on reading comprehension (0.41) for Sound Partners but the developer indicated no time targeted towards this literacy domain. To be consistent with our other analyses and to avoid "double-counting" costs attributed to different domains, we would not calculate a cost-effectiveness ratio for this domain. However, because these effects may be due to time spent on precursor skills (phonics and fluency) that have demonstrated causal relations with reading comprehension (NICHD, 2000), we could arguably calculate a cost-effectiveness ratio for reading comprehension by assigning total program costs to the relevant effect size, instead of presenting the two cost-effectiveness ratios for the precursor domains. The resulting cost-effectiveness ratio would be $1,929 per standard deviation increase in reading comprehension. Table 9 only shows the analysis that is consistent with our approach across other programs.

Vadasy and Sanders' evaluation compared the delivery of Sound Partners in one-on-one situations to dyad (one-on-two) groupings. The two modes of delivery resulted in very similar impacts on student reading outcomes but incurred substantially different costs per student. Because the effect size differences between individual and dyad modes are not significant, we use the same effect sizes (alphabetics 0.34, fluency 0.48, and comprehension 0.41) to calculate cost-effectiveness ratios for each configuration. According to the program developer, the training, facilities and all other resources used were the same for individual and dyad tutoring. Some tutors taught in both configurations. The only differences in delivery were the number of students instructed at once and a small difference in average total number of program hours per student. We split total program costs between the individual and dyad students according to the tutoring hours reported. As a result, 58% of the total costs are attributed to the individually tutored students and 42% to the students tutored in dyads. The average costs per student are $1,133 for the individually tutored students and $555 for the dyads, compared with $791 for the overall program.

As shown in Table 9, while the overall (Tutoring vs. Control) cost-effectiveness ratio for alphabetics is $2,093 per standard deviation increase in alphabetics skills, for individual tutoring this increases to $2,999 and for dyads it drops to $1,470. For the fluency domain, the overall cost-effectiveness ratio is $165 per standard deviation increase in fluency skills. For the individually tutored students this rises to $236, while for the dyads it drops to $116 per standard deviation increase in fluency skills. As the program appears equally effective in both configurations while costs drop significantly for the dyad situation, the program is clearly more cost-effective when delivered to dyads.

Table 9
Cost-effectiveness of Sound Partners

Implementation condition and literacy domain	Effect size	Significance	% of costs attributed to this domain	Average cost attributed to this domain	Cost per unit increase in effect size
Tutoring vs. control					
Alphabetics	0.34	ns	90%	$712	$2,093
Fluency	0.48	sig	10%	$79	$165
Comprehension	0.41	sig	na	na	na
Sensitivity analysis:					
i) Individual tutoring vs. control					
Alphabetics	0.34	ns	90%	$1,020	$2,999
Fluency	0.48	sig	10%	$113	$236
Comprehension	0.41	sig	na	na	na
ii) Dyad tutoring vs. control					
Alphabetics	0.34	ns	90%	$500	$1,470
Fluency	0.48	sig	10%	$56	$116
Comprehension	0.41	sig	na	na	na

Note. Costs from Table 8. Effects from Table 2. Percent delivery time targeting each domain from Table 3.

3.5 Fast ForWord Reading 1

Fast ForWord Reading 1 Program Description and Effectiveness

Fast ForWord® is a family of computer-based products which help students develop and strengthen the cognitive skills necessary for successful reading and learning. Participants spend 30 to 100 minutes a day, five days a week, for four to 16 weeks working on adaptive exercises. Fast ForWord® Language builds fundamental cognitive skills of memory, attention, processing, and sequencing in the context of key language and reading skills, including listening accuracy, phonological awareness, and language structures. Programs in the Fast ForWord® to Reading series are designed to help students acquire reading skills.

The study we identified that met WWC standards focused on one particular program: Fast ForWord Reading 1 (FFW1). The 2005 evaluation was conducted by the program developer, Scientific Learning Corporation, at three school sites, each in a different state, in the Spring semester of the 2004-2005 academic year. The study sample comprised 158 first graders and 50 second graders, all of whom were rated as slightly below-average readers by their teachers. Of the total sample, 103 students were in the treatment group and used FFW1 in a computer lab under the supervision of a trained lab monitor for an average of 43 minutes per day over 24 days. Impact of the program was measured using TOPA Letter Sounds, a measure of phonics, and TOPA Phonological Awareness. Effect sizes were 0.27 for phonological awareness and 0.20 for phonics, an average of 0.24 for the alphabetics domain. Scientific Learning Corporation kindly provided

us with site-level effectiveness data, reporting average effect sizes for alphabetics of 0.14 at Site 1, 0.41 at Site 2, and 0.28 at Site 3.[8]

FFW1 Ingredients and Costs

While the 2005 evaluation reports little detail with respect to the implementation of FFW1, we were able to interview various personnel at Scientific Learning Corporation: one of the evaluators participating in the 2005 study, an implementation manager who was able to inform us of variations in implementation at the three study sites, and several sales personnel who were able to inform us how site licensing and training operated in 2005. Table 10 shows ingredients utilized across the three sites and their associated costs.

Because the FFW1 site license, costing $50,000, covered use by an unlimited number of students at each site for an indefinite period of time, not just the students participating in the six-week evaluation, we amortized this cost over five years, based on the length of time the developer informed us that schools typically used the program. We subsequently spread this annualized cost, and the cost of the FFW1 support package ($4,500 per year), over the total number of students we assumed to have used the product over one year. We were able to estimate the total number of students using the program during the semester in which the evaluation occurred based on information from the program implementer. For the second semester we had to make assumptions regarding program use as this information was not available. In our base-case analysis, we assumed the same number of students used the program in the second semester as in the first and estimated a total cost per student for the program of $282.

It is notable that at personnel costs for the three sites together are only 26% of the total $282 per student, while the FFW1 license and support package from Scientific Learning Corporation account for 68% of all costs. Among the three sites, total cost per participant is lowest at Site 2 (see Table 10). Differences in site costs are explained by the use of literacy coaches and the scale of operation. With no literacy coach involved and the advantage of some economies of scale because fixed costs can be spread over the larger number of students served, costs per student at Site 2 were $218. A literacy coach accounted for 10% of costs at Site 1. At the third site there was also no literacy coach but fewer students were served so that costs per student were $390.

FFW1 Cost-Effectiveness and Sensitivity Analysis

Table 11 summarizes the costs, effects and cost-effectiveness of FFW1 at each of the three sites for the alphabetics domain. Based on the program developer's response to the survey regarding percent of delivery time addressing various literacy constructs (see Table 3), we assigned 50% of the program costs to the alphabetics domain. The cost-effectiveness ratios for the three individual sites vary substantially from a low of $269 per standard deviation increase in alphabetics skills at Site 2 to a high of $1,035 for Site 1 (see Table 11, Main Analysis). This suggests that the cost-effectiveness ratio for the three sites combined of $601 per standard deviation increase in alphabetics skills is not a reliable estimate of cost-effectiveness for this program. This example clearly illustrates the importance of assessing efficiency of programs by tying site-level costs to site-level effectiveness data.

8 The site-level effect sizes, obtained from the program evaluator, are calculated as Cohen's *d*, i.e., the difference between the mean outcome for the intervention group and the mean outcome for the comparison group, divided by the standard deviation of the outcome for the comparison group. While the site-level effect sizes are threfore strictly only comparable with each other, the pooled effect size of the three sites, adopted from WWC (2005) in the form of Hedges' *g*, is comparable with the effect sizes of other programs.

Table 10
Ingredients and Costs for FFW1

	Site 1 (25 students)			Site 2 (50 students)			Site 3 (28 students)			Total (103 students)		
Ingredients	Total cost	Cost per student	% of total	Total cost	Cost per student	% of total	Total cost	Cost per student	% of total	Total cost	Cost per student	% of total
Personnel total	**$2,588**	**$104**	**36%**	**$2,910**	**$58**	**27%**	**$2,046**	**$73**	**19%**	**$7,544**	**$73**	**26%**
Lab monitor	$663	$27		$1,130	$23		$607	$22		$2,400	$23	
Literacy coach	$717	$29		—	—		—	—		$717	$7	
Principal	$340	$14		$340	$7		$340	$12		$1,020	$10	
Classroom teacher	$370	$15		$765	$15		$591	$21		$1,726	$17	
Computer technician	$268	$11		$244	$5		$248	$9		$760	$7	
Parents' time	$230	$9		$431	$9		$259	$9		$920	$9	
Facilities total	**$308**	**$12**	**4%**	**$525**	**$11**	**5%**	**$316**	**$11**	**3%**	**$1,149**	**$11**	**4%**
Computer lab	$307	$12		$524	$10		$315	$11		$1,146	$11	
Training space	$1	$0		$1	$0		$1	$0		$3	$0	
Materials and equipment total	**$4,320**	**$173**	**60%**	**$7,445**	**$149**	**68%**	**$8,533**	**$305**	**78%**	**$20,298**	**$197**	**70%**
Program licenses	$2,983	$119		$5,143	$103		$5,965	$213		$14,091	$137	
SLC support	$1,229	$49		$2,120	$42		$2,459	$88		$5,808	$56	
Computer	$80	$3		$140	$3		$81	$3		$301	$3	
Headphone and Y-connector	$5	$0		$9	$0		$5	$0		$19	$0	
Projector	$2	$0		$4	$0		$2	$0		$9	$0	
Progress tracker reports	$16	$1		$26	$1		$17	$1		$59	$1	
Pen & highlighter	$4	$0		$4	$0		$4	$0		$12	$0	
Other inputs total	**$26**	**$1**	**0%**	**$33**	**$1**	**0%**	**$39**	**$1**	**0%**	**$98**	**$1**	**0%**
SLC training package	$26	$1		$33	$1		$39	$1		$98	$1	
Grand total	**$7,242**	**$290**	**100%**	**$10,913**	**$218**	**100%**	**$10,933**	**$390**	**100%**	**$29,088**	**$282**	**100%**

Note. SLC = Scientific Learning Corporation

The very small effect size obtained at Site 1 accounts for the unfavorable cost-effectiveness ratio relative to the other two sites. Inspection of the site level pre- and post-test data provided to us by Scientific Learning Corporation indicates that the students at Site 1 (both treatment and comparison group) started with higher scores on the pre-tests of phonics and phonological awareness than students at the other two sites, indicating that these students were already further in their development of these skills. Both treatment and comparison group at Site 1 improved at about the same rate, resulting in a minimal effect size. This may indicate that higher performing students perform equally well with or without FFW1, while lower performing students benefit more from FFW1 because the program emphasizes the pre-reading skills that they have not yet mastered. Site 2 benefits from both the highest effect size and the most efficient resource use, obtained by spreading fixed costs across more students. Site 3, although equally as effective as Site 2, was not as efficient because fewer students used the program throughout the year, resulting in a higher average cost per student.

As a sensitivity analysis we altered our assumption regarding the number of students using FFW1 in addition to the study participants. We repeated the base-case analysis assuming that no students used the program in the second semester. This alteration significantly affected our calculation of fixed costs per student for the FFW1 site license and support fees, representing the least efficient use of the site license and associated support services. Costs per student under these circumstances rise from $282 for the pooled sample to $479 (see Table 11, Sensitivity Test 1). The pooled cost-effectiveness ratio rises from $601 to $1,019 per standard deviation increase in alphabetics skills, reflecting less efficient use of fixed resources.

In a second sensitivity analysis (see Table 11, Sensitivity Test 2), we simulated a "most efficient" use of resources by assuming that the same number of students served in the first session of FFW1 implementation was served 3 more times over the course of the year. Under this scenario, the cost per student for the program drops to $205 and the cost-effectiveness ratio falls from $601 to $436 per standard deviation increase in alphabetics skills. Because of the high fixed costs for FFW1, the cost-effectiveness ratio is highly sensitive to the number of students being served by the program over the course of a year. We estimate that if Site 1 provided FFW1 to a total of 100 students over one year, with 25 students in each of four 6-week sessions, the school would incur total costs of around $29,000. If Site 2 also provided FFW1 over 4 sessions but to 50 students each time for a total of 200 students, we estimate that the costs incurred by the school would total only around $1,000 more, at $30,000.

Since 2005, FFW1 has been transformed from a workstation-based program to a web-based application, now called Fast ForWord Reading Level 1. As a result, the ingredients required for implementation and their associated costs have changed. Instead of a site license, schools now purchase individual licenses at $280 per student. While previously the school needed to select which Fast ForWord product to purchase, the individual license allows each student to be matched electronically with an appropriate product. Depending on the student's reading scores, as measured by the Reading Progress Indicator, the assigned product may provide accelerated or remedial instruction. Length of each Fast ForWord session has dropped from 48 minutes to 30 minutes and the program can be used anywhere the student has access to the internet, obviating the necessity for a dedicated computer laboratory. Training for adult supervisors (e.g., lab monitor or parent) can now be delivered online instead of in face-to-face sessions with Scientific Learning Corporation trainers, although both options are available.

As a final sensitivity analysis (see Table 11, Sensitivity Test 3), we model a scenario in which the web-based application of Fast ForWord Reading Level 1 is used in the same delivery context as the evaluated version of the program (FFW1), i.e., while we modify certain ingredients or their prices to reflect the current

Fast ForWord products, we keep constant the number of students served and the duration of program use. The decrease in cost of the license, availability of online training, shorter sessions, and computerization of the screening process result in a significantly lower average cost per student of $97 compared with $282 in our original workstation-based scenario. While there is no current effectiveness data to provide evidence regarding the impact of the web-based version of Fast ForWord Reading Level 1, if the effect size for the alphabetics domain remained at 0.24, the cost-effectiveness ratio would improve from $1,019 to $205 per standard deviation increase in alphabetics skills.

Table 11

Cost-effectiveness of FFW1 Across 3 Sites for Alphabetics

	Site 1	Site 2	Site 3	All three sites
# of participants in the session of interest	25	50	28	**103**
% of costs attributed to alphabetics	50%	50%	50%	**50%**
Effect size for alphabetics*	0.14	0.41	0.28	**0.24**
Main analysis				
Total program cost for one session	$7,242	$10,913	$10,933	**$29,088**
Average cost per participant	$290	$218	$390	**$282**
Average costs attributed to alphabetics	$145	$109	$195	**$141**
Cost per unit increase in effect size for alphabetics	**$1,035**	**$269**	**$697**	**$601**
i) Sensitivity test 1: most conservative scenario[a]				
Total program cost for one session	$11,534	$18,302	$19,503	**$49,339**
Average cost per participant	$461	$366	$697	**$479**
Average costs attributed to alphabetics	$231	$183	$348	**$240**
Cost per unit increase in effect size for alphabetics	**$1,648**	**$452**	**$1,244**	**$1,019**
ii) Sensitivity test 2: most efficient scenario[b]				
Total program cost for one session	$7,242	$7,489	$6,352	**$21,083**
Average cost per participant	$290	$150	$227	**$205**
Average costs attributed to alphabetics	$145	$75	$113	**$102**
Cost per unit increase in effect size for alphabetics	**$1,035**	**$185**	**$405**	**$436**

(continued on next page)

Table 11

Cost-effectiveness of FFW1 Across 3 Sites for Alphabetics *(continued)*

	Site 1	Site 2	Site 3	All three sites
iii) Sensitivity test 3: current price estimate				
Total program cost for one session	$3,269	$4,189	$2,484	**$9,943**
Average cost per participant	$131	$84	$89	**$97**
Average costs attributed to alphabetics	$65	$42	$44	**$48**
Cost per unit increase in effect size for alphabetics	**$467**	**$103**	**$158**	**$205**

Note. Site-level effect sizes obtained from Scientific Learning Corporation. Effect sizes for "All three sites" from Table 2. Costs from Table 10. Percent time attributed to alphabetics from Table 3. Note that cost per unit increase in effect size may not appear to precisely reflect effect sizes and costs reported in this table due to rounding. *Site-level effect sizes obtained from Scientific Learning Corporation are calculated as Cohen's *d*, while the pooled effect size from WWC (2007a) for "All three sites" is calculated as Hedges' *g*.

a Assumptions for Sensitivity Test 1:
Site 1: 2 x 25-student sessions in Spring and Fall 2005, 50 students in total;
Site 2: 1.16 x 50-student sessions in Spring and Fall 2005, 58 students in total;
Site 3: 1 x 28-student sessions in Spring and Fall 2005, 28 students in total.

b Assumptions for Sensitivity Test 2:
Site 1: 4 x 25-student sessions in Spring and Fall 2005, 100 students in total (same as base case);
Site 2: 4 x 50-student sessions in Spring and Fall 2005, 200 students in total;
Site 3: 4 x 28-student sessions in Spring and Fall 2005, 112 students in total.

The various sensitivity analyses we present for FFW1 demonstrate that for computer-based programs with high fixed costs, the average cost per student is extremely sensitive to scale. As effectiveness does not appear to be related to scale based on the limited data available, reducing the average cost per student by spreading fixed costs over a large number of participants significantly improves cost-effectiveness of the program. The analysis of current costs of the web-based version of the program shows that, as with many technology products, costs have fallen dramatically over time. If effectiveness remains unchanged, cost-effectiveness will also have improved.

3.6 Reading Recovery

Reading Recovery Program Description and Effectiveness

Reading Recovery is a short-term tutoring intervention intended to serve the lowest-achieving first-grade students and bring them to grade level in reading by the end of the intervention period. The goals of Reading Recovery are to promote literacy skills school-wide, reduce the number of first-grade students who are struggling to read, and prevent long-term reading difficulties. Reading Recovery supplements classroom teaching with one-to-one tutoring sessions, generally conducted as pull-out sessions during the school day. Tutoring, which is conducted by trained Reading Recovery teachers, takes place daily for 30 minutes over 12 to 20 weeks. In addition to working with individual students, the Reading Recovery teacher assumes classroom teaching responsibilities. The program requires extensive teacher training and teacher support from

a trained Teacher Leader. Intensive teacher training facilitates the broader goal of Reading Recovery which is to increase the quality of teaching for all students in all subjects.

Schwartz (2005) evaluated Reading Recovery as implemented in the field with 94 at-risk first-grade students distributed across 47 Reading Recovery teachers in elementary schools in 14 states. Several measures were used to assess program impact for 74 of these students on various early literacy outcomes. Effect sizes for literacy constructs measured in the alphabetics domain were: phonics 1.37, phonemic awareness 0.44 (ns), print awareness 1.04, and letter knowledge 0.23 (ns), for an average of 0.70 in this domain. Additionally WWC (2008) reports an effect size for fluency of 1.71, and for reading comprehension of 0.14 (ns). Other studies of Reading Recovery show variability in the effect sizes for these same literacy domains. For example, Pinnell, DeFord, and Lyons (1988), find an effect size of 0.35 (ns) for alphabetics and 0.56 for comprehension, while Iverson and Tunmer (1993), find an effect size of 1.94 for alphabetics.

Reading Recovery Ingredients and Costs

As we were not able to obtain ingredients specifically associated with the evaluated implementation, we modeled an "average" implementation of Reading Recovery based on information obtained through interviews and e-mail exchanges with the evaluator and others involved in the development and administration of the program. Reading Recovery is organized by "site", each of which consists of a number of schools in one or more districts. On average, each site comprises 14.5 schools in which 23 Reading Recovery teachers, supported by 1.18 Teachers Leaders, serve 187 first grade students per year. The Reading Recovery teachers receive a year of training from Teacher Leaders and ongoing professional development sessions during their tenure. They work intensively with four Reading Recovery students each half-year, in addition to their other teaching activities. The Teacher Leader receives a year of intensive training and ongoing support from university-level trainers. Each site maintains at least one special training facility with a one-way glass room where students can be observed as they work with a Reading Recovery teacher during training sessions. Table 12 shows ingredients and associated costs for a national average site implementation of Reading Recovery over one year. Total costs amount to approximately $775,000 per annum for the average Reading Recovery site. Of the $4,144 cost per student, 93% is attributable to personnel, mostly the Reading Recovery teachers who spend around 40% of their teaching day serving a small number of Reading Recovery students.[9]

While we were not able to interview personnel implementing the evaluated implementation of Reading Recovery, we were able to interview two current Teacher Leaders from different sites. This allowed us to cost out two alternative implementations of Reading Recovery. Results are also shown in Table 12. Both Teacher Leaders work at sites that employ more Teacher Leaders and Reading Recovery teachers than the national average: two Teacher Leaders and 65 teachers at Site 1 and three Teacher Leaders and 77 teachers at Site 2. Both sites operate three training facilities. While the national average site serves 187 students over one year, Site 1 serves around 585 first graders across 42 schools, and Site 2 serves around 693 first graders across 66 schools. Despite the increase in scale, the costs per student do not fall significantly from our $4,144 national average estimate, reflecting the fact that most of the program costs are variable, with 74% attributable to the Reading Recovery teacher. Site 1 total program costs are approximately $2,300,000, or $3,951 per student, while Site 2 total program costs are approximately $2,600,000, or $3,818 per student.

9 Our cost per student estimate of $4,144 is significantly lower than that estimated by Simon (2011): $6,631 at a 3.5% discount rate. However, our program cost per school ($53,270) matched very closely with Simon's ($53,059), despite several differences in pricing assumptions. The difference in cost per student arises because Simon assumes one Reading Recovery teacher per school, serving 8 students, while our analysis uses the national average statistic for Reading Recovery of 1.6 teachers and 12.9 students per school.

Table 12

Reading Recovery (RR) Ingredients List and Costs for a National Average Site and Two Currently Operating Sites

Ingredient	National average site			Site 1	Site 2
	Total cost	Cost per student	% of total	Total cost	Total cost
Personnel total	**$723,329**	**$3,868**	**93 %**	**$2,193,429**	**$2,483,593**
Principal	$7,398	$40		$21,430	$25,001
Classroom teacher	$9,648	$52		$27,946	$32,603
Site coordinator	$5,323	$28		$5,323	$6,143
RR Teacher	$576,059	$3,081		$1,802,296	$1,928,545
Other school staff	$5,674	$30		$41,091	$57,527
Teacher Leader pay for ongoing PD	$1,812	$10		$3,071	$4,607
RR Teacher pay for ongoing PD	$18,002	$96		$56,322	$60,267
Training for other school teachers	$1,516	$8		$4,392	$5,124
Teacher Leader	$62,816	$336		$170,348	$255,522
Teacher Leader trainee period salary	$11,282	$60		$12,581	$28,684
RR Teacher training time pay	$22,620	$121		$45,179	$75,727
Parent volunteers for ongoing PD	$173	$1		$518	$518
Training time school administrators	$810	$4		$2,346	$2,737
Parent volunteers for 1st year training	$195	$1		$585	$585
Facilities total	**$10,735**	**$58**	**1%**	**$27,306**	**$29,522**
One-way glass room	$128	$1		$246	$385
Room for RRT training	$205	$1		$392	$614
Room for RRT ongoing PD	$162	$1		$162	$162
Space for RR lessons	$8,388	$45		$24,310	$28,361
TL office space at district	$1,852	$10		$2,197	$0
Materials and equipment total	**$14,781**	**$79**	**2%**	**$43,340**	**$50,642**
Books, testing and other materials	$12,665	$68		$36,684	$42,798
Teacher Leader office computer	$8	$0		$86	$65
RR Teacher school computer	$83	$0		$234	$277
Paper reports	$2,020	$11		$6,318	$7,484
Projector for training	$6	$0		$18	$18
Other inputs total	**$26,139**	**$140**	**3%**	**$47,031**	**$82,227**
Teacher Leader training fees	$2,162	$12		$2,411	$5,497
University trainer site visits	$180	$1		$201	$458
RR teacher training fees	$16,961	$91		$30,610	$56,783
Transport: TL visits to RRT in schools	$460	$2		$1,300	$2,750
Transport: TL visits to other sites	$40	$0		$0	$101

Ingredient	National average site			Site 1	Site 2
	Total cost	Cost per student	% of total	Total cost	Total cost
Transport: students to training sites	$181	$1		$346	$475
RRCNA conference for TL	$1,416	$8		$2,400	$3,600
TL Institute for TL	$1,416	$8		$2,400	$3,600
On campus housing for TL trainees	$117	$1		$132	$297
Technical support fee	$885	$5		$1,500	$2,250
IDEC Data Evaluation fee	$1,485	$8		$3,275	$3,815
RRCNA membership fee	$675	$4		$1,820	$2,197
Transport students to ongoing PD	$160	$1		$480	$403
Student incentives for RRT training	$0	$0		$141	$0
Student incentives for ongoing PD	$0	$0		$15	$0
Grand total	**$774,985**	**$4,144**	**100%**	**$2,311,106**	**$2,645,983**
Number of students served	*187*			*585*	*693*
Average cost per student	*$4,144*			*$3,951*	*$3,818*

Note. TL = Teacher Leader; RRT = Reading Recovery Teacher; PD =professional development; RRCNA= Reading Recovery Council of North America; IDEC = International Data Evaluation Center.

Reading Recovery Cost-Effectiveness Analysis and Sensitivity

The impact of Reading Recovery on several measures of early literacy was assessed by Schwartz (2005). Because we were not able to obtain ingredients data specifically for the evaluated implementations, we made the less-than-ideal assumption that the effect size observed in this study would be observed in the "average" implementation of Reading Recovery. This may or may not hold true. Schwartz's evaluation was more of a field study than the other evaluations we reviewed in that it appears to have involved sites already using the program, with no additional training provided for study purposes and no apparent fidelity of implementation checks beyond what might already be routine in typical program delivery. Therefore we have some confidence that the average implementation can expect similar impact on literacy outcomes. In addition, the program is highly stylized with standard procedures so that site-level variations in implementation should be minimal.

Table 13 shows the cost-effectiveness ratios for Reading Recovery in the alphabetics and fluency domains, assuming that the program costs can be split equally among the four domains of alphabetics, vocabulary, fluency, and reading comprehension. One domain, vocabulary, was not measured in Schwartz's 2005 evaluation. However, we assume the program targets this domain based on the developers' assertion that the program covers all aspects of early literacy. Additionally, program materials, such as the Reading Recovery Council of North America website, explicitly mention vocabulary. We therefore attribute a portion of the program costs towards it. No cost-effectiveness ratio is shown for reading comprehension because the effect size was not significant. The cost-effectiveness ratios are $1,480 per standard deviation increase in

alphabetics skills, and $606 per standard deviation increase in fluency skills. The ratio is more favorable for fluency because the program impact on fluency was substantially higher than on alphabetics. We note that our analysis cannot capture any improvements in the literacy or other outcomes of non-Reading Recovery students served by the highly trained Reading Recovery teachers during their regular classroom teaching activities as these impacts were not measured.

If the same effect sizes observed by Schwartz (2005) and applied to the national average site could also be applied to Sites 1 and 2, the cost-effectiveness ratios would fall slightly. In the alphabetics domain, the cost-effectiveness ratios would be $1,411 per standard deviation increase in alphabetics skills for Site 1 and $1,363 for Site 2, compared with $1,480 for the national average site. In the fluency domain, the cost-effectiveness ratio would be $578 per standard deviation increase in fluency skills for Site 1, and $558 for Site 2, compared with $606 for the national average site.

Because Reading Recovery teachers are the most significant cost of the program, costs per student and cost-effectiveness ratios will only drop markedly if each teacher serves more students. While recognizing that the program philosophy requires individualization of instruction, we conducted a sensitivity analysis in which each Reading Recovery teacher served 10 students per year and found that the national average cost per student of $4,144 fell to $3,372 per student. Applying the same 0.7 effect size for the alphabetics domain as observed by Schwartz (2005), the cost-effectiveness ratio would fall to $1,204 per standard deviation increase in alphabetics. However, we caution that there is no guarantee that different implementations will result in similar effect sizes to those observed in the evaluated implementation.

We also conducted a sensitivity analysis in which we attributed only 50% of the costs of Reading Recovery teacher training to the Reading Recovery program, under the assumption that the other half of the costs can be attributed to the teacher's regular teaching activities. The cost per student decreased from $4,144 to $3,988. As a result, the cost-effectiveness ratio dropped slightly to $1,424 per standard deviation increase in alphabetics skills from $1,480.

Finally, we provide a sensitivity analysis in which we assume that only 10% of program delivery time addresses alphabetics, instead of $25%. In this case, the cost-effectiveness ratio would fall to $592 per standard deviation increase in alphabetics skills.

Table 13
Cost-effectiveness Ratios for Reading Recovery (RR)

Literacy domain	Effect size assumed	Cost per unit increase in effect size		
		National average site	Site 1	Site 2
Alphabetics	0.70	$1,480	$1,411	$1,364
Fluency	1.71	$606	$578	$558
Sensitivity analysis:				
i) Each RR teacher serves 10 students per year				
Alphabetics	0.70	$1,204		
Fluency	1.71	$493		

Table 13

Cost-effectiveness Ratios for Reading Recovery (RR) ***(continued)***

Literacy domain	Effect size assumed	Cost per unit increase in effect size		
		National average site	Site 1	Site 2
ii) 50% of RR teacher training cost attributed to RR program				
Alphabetics	0.70	$1,424		
Fluency	1.71	$583		
iii) 10% of costs attributed to alphabetics				
Alphabetics	0.70	$592		

Note. 25% of total program costs are attributed to each of the four literacy domains except in sensitivity analysis (iii). Total costs of $4,144 (national average site), $3,951 (Site 1), and $3,818 (Site 2) per student are from Table 12. Effect sizes are from Table 2.

3.7 Corrective Reading and Wilson Reading System

Students who are still struggling to read by the third grade are likely to be those with more entrenched reading difficulties than younger struggling students and therefore harder to remediate than kindergarten or first grade students. Torgesen et al. (2007), among many others, report that reading difficulties for third-grade students range from those who struggle with phonemic decoding to those who read fluently but have difficulty comprehending what they read. Not only must third-grade struggling readers catch up on basic reading skills in the alphabetics domain that most students have grasped by the end of first grade, but they are behind their peers in fluency, vocabulary and reading comprehension overall. As such, it is likely that interventions to help these students will need to be more intensive and of longer duration (as well as different in curricular content), than interventions for younger struggling students. Therefore, our expectation for programs that serve struggling readers later in elementary school is that costs will be higher and effect sizes lower than for programs serving younger students. The resulting cost-effectiveness ratios should therefore be substantially higher, i.e., less favorable.

Two reading programs serving third grade students met our selection criteria and are listed by WWC as showing potentially positive effects in the alphabetics domain: Wilson Reading System and Corrective Reading. These two are presented together in our final cost-effectiveness comparison because they were both evaluated in the same study by Torgesen et al. (2006) in 2003-2004. The Torgesen et al. study provides the ideal scenario for a cost-effectiveness analysis because several conditions desirable for program comparability are met: there is consistency in grade level and reading ability of the students served, duration of intervention, and outcome measures used to assess effectiveness. However, we note that both programs were modified substantially to fit with the evaluation study design. Consequently, while our analysis provides a text-book demonstration of a cost-effectiveness comparison, it cannot reflect the true cost-effectiveness of the programs as they are normally implemented. Additionally, we note below a number of issues regarding the Torgesen et al. study which highlight a problem often associated with rigorously designed randomized control trials: while findings may have high internal validity, they are not always a realistic reflection of how effective the interventions might be in typical school implementations.

Both programs were implemented with third and fifth grade struggling readers in different schools in the Pittsburgh, Pennsylvania area for 28 weeks. We focus only on the third graders. In some instances students in the study missed part of their regular reading time to attend the Corrective Reading and Wilson Reading System sessions despite the fact that each developer recommends their program as a supplement to classroom instruction, not as a partial substitute. This may have negatively affected the potential effectiveness of the programs. On the other hand, greater than usual ongoing training throughout the year for the carefully selected teachers delivering the programs, and the intense fidelity of implementation measures may have biased effectiveness upwards.

Corrective Reading

Corrective Reading Program Description and Effectiveness

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, each of which is split into four levels. The developer recommends 45-minute sessions four to five times per week over a period of two to three years. The entire decoding strand consists of 320 lessons that would take 64 weeks (almost two academic years) to deliver at five sessions per week, totaling 240 hours of instruction. This can be accelerated by delivering two sessions per day when feasible. The comprehension strand consists of 330 lessons that could take almost another two years to deliver.

For the purposes of the Torgesen et al. (2006) evaluation, only the decoding strand was used and instruction was delivered by ten trained teachers to groups of three third grade students, one hour per day, five days per week over 28 weeks. On average, treatment students in the study received 90 hours of instruction, far short of the 240 hours intended by the program developer. Only the middle two of the four decoding levels (B1 and B2) were covered by the majority of the students, and instructional emphasis was greater on alphabetics outcomes than fluency or comprehension. Specifically, according to a McGraw-Hill Product Information specialist who was familiar with the study, 83% of delivery time on average was targeted at alphabetics (compared with 65% of regular program delivery) and 17% was targeted at fluency (compared with 20% normally). In addition, regular program implementations would spend around 15% of delivery time addressing comprehension skills. Teachers in the study were instructed to accelerate instruction by covering more than one 45-minute lesson in the 60-minute period each day. This resulted in omissions of material and partial lessons being delivered.

The impact of Corrective Reading was measured by Torgesen et al. (2006) using four measures of alphabetics, one measure of fluency, and two measures of comprehension. In the alphabetics domain, the overall average effect size for the four measures was 0.22. In the fluency domain, a statistically significant effect size of 0.27 was observed in spite of only 17% of delivery time spent addressing this domain. In the comprehension domain, a positive but not statistically significant effect size of 0.17 was observed. To set these effect sizes in perspective, Hill et al. (2007) report that students progressing through third grade to fourth grade show an average annual reading gain of 0.36 standard deviation.

Corrective Reading Ingredients and Costs

With the assistance of the Institute for Education Sciences, which commissioned the Torgesen et al. evaluation, we were able to obtain identities of the schools, teachers and trainers participating in the study of Corrective Reading and Wilson Reading System. Unfortunately, but not surprisingly given that the study was conducted ten years ago, most teachers had moved from the school districts participating in the study. However, we were able to interview one of the four remaining Corrective Reading teachers out of the original ten who participated in the evaluation. This teacher was able to provide us with substantial detail regarding program ingredients and implementation, beyond what was available in the already extensive evaluation report. In addition, two individuals at the McGraw-Hill School Education Product Information Center were able to provide additional details about the study implementation and how the modified version of Corrective Reading used in the evaluation differed from the original version.

Table 14 lists the ingredients required to implement Corrective Reading for one teacher with a group of three students. This is based on information from our teacher interviewee as well as the Torgesen et al. (2006) evaluation report, but we note that the experience of this teacher differed from some of the other teachers in the program. For example, while this teacher taught three groups of students, others taught four. In some instances our interviewee's report differed from what was described in the evaluation report and from information that we obtained from our Wilson Language Training interviewees regarding the same study. Where it appeared that the differences were due to difficulty in remembering details from ten years ago, we abided by the evaluation report, but other differences appear to reflect genuine deviations in implementation from the average reported.

Our interviewee taught one third grade group and one fifth grade group that were part of the study and an additional fourth grade group which was not part of the study. We split the costs of the program across the three grade levels served, therefore assigning 33% of costs to each grade level. Based on the experience of our interviewee, we assume that the teacher worked full-time to deliver the program to three groups of three students daily throughout the study period, one group in each of third, fourth and fifth grades. The total cost of the program per student is $10,108 with the Corrective Reading teacher's salary accounting for 87% of the costs. For teachers serving four groups of students per day, costs per student would be lower as we later illustrate in a sensitivity analysis.

Table 14

Ingredients and Costs of Corrective Reading (CR)

Ingredients total	Total cost (3 students)	Cost per student	% of total costs
Personnel total	**$28,625**	**$9,542**	**94%**
Classroom Teacher	$18	$6	
CR Teacher	$26,371	$8,790	
Substitute Teachers	$766	$255	
CR Teachers – selection and training	$345	$115	
Local coordinators	$249	$83	
Trainers – training	$136	$45	
Trainers – implementation fidelity and support	$572	$191	
Testers to administer screening tests	$79	$26	
Testers – training time	$4	$1	
Tester training – trainers	$17	$6	
Substitute Teachers – training time	$24	$8	
Parents for conferences	$43	$14	
Facilities total	**$901**	**$300**	**3%**
Classroom	$900	$300	
Room for training	$1	$0	
Tester training – facilities	$0	$0	
Materials and equipment total	**$406**	**$135**	**1%**
Lesson materials – durable	$268	$89	
Lesson materials – consumable	$116	$39	
Screening tests	$14	$5	
Camcorder	$0	$0	
Cassette tape	$2	$1	
Postage for video	$3	$1	
Training	$1	$0	
Star posters	$1	$0	
Other inputs total	**$393**	**$131**	**1%**
Travel to Training – CR teachers	$70	$23	
Trainer Travel – per diems	$80	$27	
Trainer Travel – hotel stays	$122	$41	
Trainer travel – flights	$121	$40	
Grand total	**$30,325**	**$10,108**	**100%**

Corrective Reading Cost-effectiveness and Sensitivity Analysis

Table 15 summarizes the costs, effects and cost-effectiveness of Corrective Reading for alphabetics and fluency. While impact on comprehension was also measured, we do not calculate a cost-effectiveness ratio for this domain because the modified version of the program used in the Torgesen et al. (2006) study did not target this domain and the effect size is not statistically significant. For alphabetics, the cost-effectiveness ratio is $38,135 per standard deviation increase in alphabetic skills. For fluency the ratio is $6,364 per standard deviation increase in fluency skills. As expected, these numbers are much higher than for the programs targeting younger students, perhaps reflecting the increasing difficulty of addressing reading deficiencies as students get older.

As the most significant cost of the program is the Corrective Reading teacher's salary, cost-effectiveness should improve if the teacher is able to teach more children each year, either using this program, or performing other teaching activities in the school. Our costs above are based on the teacher spending 33% of her time with three third grade students. If the teacher instead spent only 25% of her time with these students, costs would fall to $7,771 per student and the cost-effectiveness ratio for alphabetics would drop from $38,135 per standard deviation increase in alphabetics skills to $29,318. The cost-effectiveness ratio for fluency would drop from $6,364 to $4,893 per standard deviation increase in fluency skills (See Sensitivity analysis (i) in Table 15).

Another reason for high personnel costs associated with Corrective Reading is that the teachers in the Torgesen et al. (2006) study had an unusually high level of teaching experience, averaging just over 15 years. We expect that Corrective Reading could be implemented with less experienced, and thereby less costly, teachers, although we acknowledge that effectiveness may also be affected. Studies of teacher experience suggest that skill improvement is concentrated in the first few years of teaching and then tapers off (Harris & Sass, 2011). If the experience level of the Corrective Reading teacher were five years rather than 15, and the teacher taught four groups of three students per day, the cost per student would drop further to $6,332. If the impact on literacy domains remained the same as observed in the Torgesen et al. study, the cost-effectiveness ratio for alphabetics would drop to $23,889 per standard deviation increase in alphabetics skills. The cost-effectiveness ratio for fluency would drop to $3,987 per standard deviation increase in fluency skills.

Table 15

Cost-effectiveness Ratios for Corrective Reading

Literacy domain	Effect size	% of total costs	Cost per student attributed to this domain	Cost per unit increase in effect size
Alphabetics	0.22	83%	$8,390	$38,135
Fluency	0.27	17%	$1,718	$6,364
Comprehension	0.17 ns	0%	na	na
Sensitivity analysis:				
(i) Four groups per teacher				
Alphabetics	0.22*	83%	$6,450	$29,318
Fluency	0.27*	17%	$1,321	$4,893
Comprehension	0.17 ns*	0%	na	na
(ii) Four groups per teacher, 5 years experience				
Alphabetics	0.22*	83%	$5,256	$23,889
Fluency	0.27*	17%	$1,076	$3,987
Comprehension	0.17 ns*	0%	na	na

Note. ns = not significant; na =not applicable. Costs are from Table 14. Percent time attributed to each domain is based on the program developer's indications of percent of program delivery time addressing constructs in this domain, see Table 3. *Effect size was not measured under these conditions but we assume effect size would not change with these modifications to program implementation.

Wilson Reading System

Wilson Reading System Program Description and Effectiveness

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 small groups of students (one to six), 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.

For the purposes of the Torgesen et al. study, the Wilson Reading System program was delivered to the third grade students over 28 weeks without the vocabulary and comprehension components, and sessions were delivered five times per week for no longer than 60 minutes each. During the study period, students progressed from Step 1 to between Steps 4 and 6 of the curriculum. According to the Wilson Language Director of Research and Evaluation, and the trainer who participated in the study, delivery time during the study, on average, addressed literacy domains as follows: alphabetics 66%, fluency 2%, comprehension 4%,

and writing 28%. In a regular implementation of the program relatively less time would be spent on alphabetics (46%), and writing (13%), and more on fluency (11%), comprehension (19%), and vocabulary (11%).

The impact of Wilson Reading System was assessed by Torgesen et al. with four measures of alphabetics, one measure of fluency, and two measures of comprehension. These were the same measures used for Corrective Reading. In the alphabetics domain, the overall average effect size across the four measures was 0.33. In the fluency domain, a positive, but not statistically significant, effect size of 0.15 was observed, and for comprehension, a positive but not statistically significant effect size of 0.17 was observed.

Wilson Reading System Ingredients and Costs

Of the nine teachers who participated in the Torgesen et al. (2006) evaluation, we could only locate three, and none were available or willing to be interviewed at the time of our data collection. However, we were able to interview the Wilson Language trainer who trained all nine teachers during the study period and visited them regularly at the study sites. In addition, the individual at Wilson Language Training who served as the administrative contact for the Torgesen et al. study participated in the interview, as did the Wilson Language Director for Research and Evaluation. In addition to details regarding the program implementation and ingredients required, our interviewees informed us how the modified version of Wilson Reading System used in the evaluation differed from the original version.

In contrast to Corrective Reading where we present the costs for one teacher that were attributable to a single third grade group, for Wilson Reading System we present the costs for all nine teachers trained by our Wilson Language interviewee that were attributable to the entire third grade cohort (53 students). While each approach reflects a different perspective, both aim to capture the total costs per student so that the cost per student estimates are comparable. There was only one discrepancy between the Wilson Language trainer and the Corrective Reading teacher in reporting ingredients that were common across the programs. This concerned the role of the local study coordinator, but given that the cost of this person accounted for 1% of total costs, the discrepancy is not significant.

Table 16 shows the ingredients and associated costs for Wilson Reading System. Total costs per student are $6,696, 78% of which are for the Wilson Reading System teachers. Note that we split the salaries for these teachers across all students served by the program. This includes the third and fifth graders included in the study, and we also attributed a small portion of the teachers' salary to the fourth graders who benefited from the program during a preliminary practice period.

Table 16

Ingredients and Costs for Wilson Reading System (WRS)

Ingredients	Total cost (53 students)	Cost per student	% of total costs
Personnel total	**$321,326**	**$6,063**	**91%**
Classroom teacher	$0	$0	
Substitute teachers	$18,184	$343	
Local coordinators	$1,810	$34	
Trainers – training	$2,017	$38	
Trainers – implementation fidelity and support	$13,684	$258	
Testers to administer screening tests	$1,401	$26	
Testers – training time	$64	$1	
Tester training - trainers	$249	$5	
Original substitute teacher – training time	$351	$7	
Additional substitute teachers – training time	$387	$7	
WRS teachers – after-school meetings	$2,274	$43	
WRS teachers	$276,410	$5,215	
WRS teachers – selection and training	$4,496	$85	
Facilities total	**$24,688**	**$466**	**7%**
Classroom	$24,665	$465	
Room for training	$19	$0	
Tester training – facilities	$1	$0	
Additional substitute teachers – training facilities	$3	$0	
Materials and equipment total	**$3,449**	**$65**	**1%**
Lesson materials – durable	$333	$6	
Lesson materials – consumable	$2,308	$44	
Classroom materials – per student	$245	$5	
Classroom materials – per teacher	$73	$1	
Classroom materials – per teacher, durable	$175	$3	
Screening tests	$256	$5	
Camcorder	$0	$0	
Cassette tape	$14	$0	
Postage for video	$26	$0	
Training videos	$8	$0	
Computer and internet time for self-study	$12	$0	
Other inputs total	**$5,451**	**$103**	**2%**
Travel to initial training – WRS teachers	$305	$6	
Travel to ongoing training – WRS teachers	$405	$8	
Trainer travel – per diems, hotel, flights	$4,740	$89	
Grand total	**$354,914**	**$6,696**	**100%**

Wilson Reading System Cost-effectiveness and Sensitivity Analysis

For the alphabetics domain, the cost-effectiveness ratio for Wilson Reading System is $13,392 per standard deviation increase in alphabetics skills, as shown in Table 17. We do not show cost-effectiveness ratios for fluency or comprehension because the measured impacts were not statistically significant. As for Corrective Reading, these numbers are much higher than for the programs targeting younger students, perhaps reflecting the greater challenge of remediating reading problems after first grade.

In our baseline estimate above, we calculate the costs of Wilson Reading System assuming that each teacher serves four groups of three students per day for one hour each. However, Wilson Reading System can be used in groups up to six students in size. If we alter our baseline scenario to model each teacher serving four groups of six students, the cost per student falls from $6,696 to $3,328. This analysis illustrates how the most significant cost of the program can be minimized, although we cannot be sure that the same level of effectiveness would be maintained as observed in the smaller groups. If we assume the same level of effectiveness for the alphabetics domain, the cost-effectiveness ratio would fall from to $13,392 to $6,656 per standard deviation increase in alphabetics skills (See Table 17, Sensitivity Test (i)).

If Wilson Reading System were implemented by teachers with an average of five years of teaching experience, as opposed to the nine years of experience reported for the Torgesen et al. (2006) teachers, the cost per student would drop to $6,188 and the cost-effectiveness ratio would fall from to $13,392 to $12,376 per standard deviation increase in alphabetics skills (See Table 17, Sensitivity Test (ii)).

Table 17
Cost-effectiveness Ratios for Wilson Reading System

Literacy domain	Effect size	% of total costs	Cost per student attributed to this domain	Cost per unit increase in effect size
Alphabetics	0.33	66%	$4,420	$13,392
Fluency	0.15 ns	2%	$134	na
Comprehension	0.17 ns	4%	$268	na
Sensitivity analysis:				
(i) Classes of six students				
Alphabetics	0.33*	66%	$2,196	$6,656
Fluency	0.15 ns*	2%	$67	na
Comprehension	0.17 ns*	4%	$133	na
(ii) Teachers with 5 years of experience				
Alphabetics	0.33*	66%	$4,084	$12,376
Fluency	0.15 ns*	2%	$124	na
Comprehension	0.17 ns*	4%	$248	na

Note. ns = not significant; na =not applicable. Costs are from Table 16. Percent time attributed to each domain is based on the program developer's indications of percent of program delivery time addressing constructs in this domain, see Table 3. *Effect size was not measured under these conditions but we assume effect size would not change with these modifications to program implementation.

4. COMPARING THE COST-EFFECTIVENESS OF EARLY LITERACY PROGRAMS SERVING THE SAME GRADE

4.1 Summary of Cost-effectiveness Results

The foregoing cost-effectiveness analyses of the seven literacy programs we investigated indicate that there is prima facie evidence that these interventions vary substantially in terms of their costs and cost-effectiveness in boosting alphabetics and other early literacy outcomes. While all early literacy interventions reviewed in this study had demonstrated positive impact on at least one measure of early literacy, costs of implementing the programs varied considerably from $27 per student for K-PALS to $10,108 per student for Corrective Reading. These differences reflect varying resource requirements, degrees of investment in teacher training, treatment duration, and dosage.

Despite the high costs of some of these programs, we note that if the interventions are successful in helping struggling readers achieve grade level reading skills, future costs of special education for program participants can be avoided. With special education expenditures for non-homebound students being 1.91 times greater, on average, than expenditures on regular education students (Chambers, Shkolnik, & Pérez, 2003) the potential savings are significant. One year of special education avoided would more than cover the cost of the even the most expensive of the programs.

We emphasize that direct comparisons among the programs should be made with caution given differences in age and reading ability of populations served, and in targeted literacy outcomes. Bearing these differences in mind, we find that the cost-effectiveness ratios for the alphabetics domain for these seven interventions vary substantially (as summarized in Table S1). The costs per standard deviation increase in alphabetics skills for the programs as evaluated are: $38 for K-PALS (at the workshop level of implementation); $570 for Stepping Stones; $601 for FFW1; $1,480 for Reading Recovery; $2,093 for Sound Partners; $13,392 for Wilson Reading System; and $38,135 for Corrective Reading. While, all other things being equal, we expect programs serving older children to appear less cost-effective than those serving younger children, the wide differences found for programs serving the same grade students is surprising. The following sections present comparisons of the programs serving students in the same grade.

4.2 Cost-effectiveness of Literacy Programs Serving Kindergarten Students

The low cost-effectiveness ratios for K-PALS at all three levels of implementation investigated primarily reflects the fact that the program does not require additional personnel beyond regular schoolteachers, except for the trainers on the initial training day. Furthermore, because it replaces some of the regular classroom instruction, there is no extra instructional time to account for during program delivery. As a result, this program appears far more cost-effective than one-to-one programs that incur additional costs beyond regular instruction. We also note that K-PALS serves all readers, unlike the other programs which serve struggling or below-average readers. However, it is unclear whether this would lead to a higher or lower cost-effectiveness ratio.

We do not compare the cost-effectiveness of K-PALS directly with that of the other two kindergarten programs we analyzed, Stepping Stones and Sound Partners, because of the difference in reading ability of students served. However, if it were possible to break out the impact on struggling readers in the K-PALS study sample, the cost-effectiveness of this mixed-ability, whole-class intervention for the subset of struggling readers could be compared with the cost-effectiveness of the more individualized and costly instruction provided by Sound Partners and Stepping Stones.

Comparing Stepping Stones and Sound Partners

We compare the Stepping Stones cost-effectiveness ratio for alphabetics with the equivalent cost-effectiveness ratio for Sound Partners because both programs serve struggling readers in kindergarten. We would expect both programs to be more expensive than K-PALS because they are delivered one-on-one or one-on-two as a supplement to classroom instruction. However, we note that Sound Partners is delivered over 18 weeks while Stepping Stones is delivered over five to six weeks. The population of students served by Stepping Stones in the study we relied on for effectiveness data were at risk for behavioral disorders, as well as being struggling readers. While for the alphabetics domain, Stepping Stones appears to be more cost-effective than Sound Partners, the program only addresses this domain, while Sound Partners also shows positive impacts on fluency and reading comprehension (see Table 2). Because Stepping Stones was not evaluated for impact on these domains, we are not able to assess overall relative efficiency of the two programs.

Table 18

Cost-effectiveness Ratios for Stepping Stones and Sound Partners for Alphabetics

Program	Effect size	% of total costs attributed to alphabetics	Cost per student attributed to alphabetics	Cost per unit increase in effect size
Stepping Stones	0.84	100%	$479	$570
Sound Partners	0.34	90%	$712	$2,093

Note. Costs for Stepping Stones are from Table 6. Costs for Sound Partners are from Table 8 and are adjusted for percent of program delivery time spent addressing alphabetics, from Table 3. Stepping Stones is a five-week program while Sound Partners extends for 18 weeks.

4.3 Cost-effectiveness of Literacy Programs Serving First Grade Students

Reading Recovery and FFW1 both serve students in first grade who are below average or struggling readers. As a result of high fixed costs, the cost-effectiveness of FFW1 appears to be highly sensitive to the number of students using the program at a site, such that it is most resource-efficient when large numbers of students use the program. Site-level differences in effectiveness suggest that the program may be most efficiently used for large numbers of students with weak reading skills. Reading Recovery represents a significant long-term commitment for multiple schools in one or more districts at once. It requires a large investment of resources for training personnel and targets only a small group of at-risk first grade students in each school, an average of 12.9 per school nationally. However, the program is a comprehensive literacy program designed to bring struggling readers back to mainstream participation.

In Table 19 we compare the cost-effectiveness ratios of FFW1 and Reading Recovery for alphabetics, but offer several caveats in making this comparison. FFW1 appears to be more cost-effective than Reading Recovery: $601 per standard deviation increase in alphabetics skills for FFW1 vs. $1,480 for Reading Recovery. We note, however, that Reading Recovery students in the evaluated implementation appear to have been weaker readers (bottom 20th percentile) than the FFW1 readers (below-average). Additionally, FFW1 served some second graders and was delivered over about six weeks compared with 12-20 weeks for Reading

Recovery. Reading Recovery outcome measures were arguably more distal to the instruction received than the outcome measure used for FFW1, and were repeated six months after the first cohort of participants completed the program, therefore offering stronger evidence of effects and their persistence.

If our assumption that 25% of program costs for Reading Recovery can be attributed to alphabetics is inaccurate and less time should be attributed to this literacy domain, then cost-effectiveness for alphabetics would improve. For example, if only 10% of program delivery time addresses alphabetics, the cost-effectiveness ratio would fall to $592 per standard deviation increase in alphabetics skills, very close to the result for FFW1.

A further limitation of the comparison between the two programs is that while the available effectiveness data allows for a comparison with respect to alphabetics outcomes, data reporting program impact on other literacy outcomes is only available for Reading Recovery. Reading Recovery shows a positive impact on fluency and on two writing outcomes (WWC, 2008), but none of these outcomes were measured for FFW1. Consequently, it is difficult to provide an overall assessment of the relative efficiency of the two programs across multiple literacy domains. If the only goal for adoption of a reading program were to improve alphabetics skills, investment in FFW1 would represent the more efficient use of resources. However, if other impacts are sought, insufficient information is available to make a full comparison between these two programs.

Table 19

Cost-effectiveness Ratios for FFW1 and Reading Recovery for Alphabetics

Program	Effect size	% of total costs attributed to alphabetics	Cost per student attributed to alphabetics	Cost per unit increase in effect size
FFW1	0.24	50%	$141	$601
Reading Recovery	0.70	25%	$1,036	$1,480

Note. Costs for FFW1 are from Table 10. Costs for Reading Recovery are from Table 12. Total costs for FFW1 are adjusted for percent of program delivery time spent addressing alphabetics from Table 3. For Reading Recovery we assumed costs were split equally across the four literacy domains, with 25% attributed to alphabetics. FFW1 is a six-week program while Reading Recovery extends for 12 to 20 weeks.

4.4 Cost-effectiveness of Literacy Programs Serving Third Grade Students

The much higher cost-effectiveness ratios obtained for the modified versions of Wilson Reading System and Corrective Reading suggest that it is more cost-effective to address reading problems as early as possible, although this may not guarantee successful remediation for the most challenged readers or address the needs of readers whose difficulties emerge after first grade (e.g., Leach, Scarborough, & Rescorla, 2003). Similarly to Reading Recovery, these programs require a significant investment in both initial and ongoing training for the instructors and the instructor salary accounts for the majority of the costs. As a result, the cost per school is highly dependent on the number of students served per instructor. While costs can be reduced by increasing the number of students per teacher, it is not clear how this might affect impact on reading outcomes. The study of Sound Partners (Vadasy & Sanders, 2008) indicated that impact on reading

outcomes for kindergarten students was not diminished by moving from one-to-one to one-to-two instruction, but it is not clear whether and how much impact might change as groups grow to three, four or more students or how this applies to other programs and different grades.

The difference in estimated cost per student between Wilson Reading System ($6,696) and Corrective Reading ($10,108), based on the Torgesen et al. (2006) study, can be attributed to several factors, some of which are likely fundamental to each program's design, and others of which are likely arbitrary, idiosyncratic features of the sample of teachers and students in the study. First, the Corrective Reading teachers had, on average, six years more teaching experience than Wilson Reading System teachers and are, in our analysis, therefore assigned higher salaries commensurate with their average level of experience. Further, the nine Wilson Reading System teachers taught 53 third grade students (about six each on average), while the ten Corrective Reading teachers taught 44 third grade students (about four and one half each on average). Finally, materials costs for Corrective Reading are about double those of Wilson Reading System, at $128 per student per year vs. $59 per student per year. However, in both cases, materials only represent approximately one percent of the total cost of the program.

Table 20 indicates that Wilson Reading System is more cost-effective than Corrective Reading for the alphabetics domain ($13,392 vs. $38,135 per standard deviation increase in alphabetics skills). This is partially due to the fact that the effect size for alphabetics is higher for Wilson Reading System than for Corrective Reading. Additionally, according to the program developers, 83% of program delivery time for Corrective Reading addressed alphabetics in the evaluated implementation while only 66% of delivery time for Wilson Reading System addressed this domain (see Table 3). As a result, a greater portion of program costs are assigned to alphabetics for Corrective Reading than for Wilson Reading System. It should also be noted that Corrective Reading shows a positive impact on fluency in addition to alphabetics. As the fluency impact for Wilson Reading System is not statistically significant, we do not calculate a cost-effectiveness ratio for Wilson Reading System in fluency that we can compare to the Corrective Reading result.

As noted above, the major reason the costs for Corrective Reading are substantially higher than those for Wilson Reading System is that, in Torgesen et al. (2006), fewer students were served by more experienced teachers in the Corrective Reading implementation compared with the Wilson Reading System implementation. However, our sensitivity analyses show that if both programs are delivered by teachers with five years of teaching experience to four groups of three students each, the costs per student are very close: $6,188 for Wilson Reading System and $6,332 for Corrective Reading. If the same effect sizes observed for alphabetics by Torgesen et al. (2006) applied in this scenario, the cost-effectiveness ratios would be $12,376 per standard deviation increase in alphabetics skills for Wilson Reading System and $23,889 for Corrective Reading.

Table 20

Cost-effectiveness Ratios for Corrective Reading and Wilson Reading System

Literacy domain	Effect size	% of total costs attributed to alphabetics	Cost per student attributed to this outcome	Cost per unit increase in effect size
Alphabetics				
Corrective Reading	0.22	83%	$8,390	$38,135
Wilson Reading System	0.33	66%	$4,420	$13,392
Fluency				
Corrective Reading	0.27	17%	$1,718	$6,364
Wilson Reading System	0.15 ns	2%	$134	na
Comprehension				
Corrective Reading	0.17 ns	0%	na	na
Wilson Reading System	0.17 ns	4%	$268	na

Note. ns = not significant; na =not applicable. Costs for Corrective Reading are from Table 14. Costs for Wilson Reading System are from Table 16. Percent time attributed to each domain is based on the program developers' indications of percent of program delivery time addressing constructs in this domain, see Table 3.

One pattern that emerges across the comparisons that we make for programs serving the same-grade students, is that in each case the program that shows an impact on other outcomes in addition to alphabetics (Sound Partners, Reading Recovery, and Corrective Reading) appears less cost-effective than the comparison program (Stepping Stones, FFW1, and Wilson Reading System respectively) when judged on the alphabetics outcome alone.

5. CONCLUSIONS

The initial goal of this research was to identify the most cost-effective ways to improve early literacy. However, our endeavors have led us to focus more on the methodological and data requirements for performing cost-effectiveness analysis. Most of these requirements would involve changes in how the effectiveness of literacy interventions is evaluated. Our economic contribution of cost analysis, which is rarely performed, depends critically on having comparable effectiveness outcomes. Differences in age and reading ability of populations served in the evaluations of literacy programs, and in measures used to assess impact, present a challenge not only for cost-effectiveness analysis, but also for policymakers who need to decide which intervention to select.

We recommend that future evaluations of early literacy programs include common outcome measures to facilitate comparability among programs. Studies in which two or more alternative programs are implemented with similar populations of students, and literacy outcomes are compared using the same measures, would greatly facilitate comparability not only of program effectiveness, but also of cost-effectiveness.

A major challenge for valuing early literacy programs is that programs often impact multiple literacy outcomes. Even if impact on all literacy domains was consistently assessed, there is no obvious way to combine the results across the four domains (alphabetics, fluency, vocabulary, and reading comprehension), a challenge made worse by the hierarchically, sequentially, and causally related nature of literacy domains. One potential solution we considered was weighting the cost-effectiveness ratios for each literacy domain by "importance of contribution" towards reading comprehension, and combining these weighted ratios to provide an overall "literacy" cost-effectiveness ratio for each program. However, because the early literacy domains are interdependent and sequential, a suitable weighting scheme is elusive.

Such weights could be assigned more subjectively by an individual decision-maker who, for example, might decide that students at his/her school need more focus on fluency than alphabetics, and therefore might weight a program's cost-effectiveness ratio for fluency by 0.7, and the cost-effectiveness ratio for alphabetics by 0.3. The combination of these weighted ratios for one program could then be compared with a similar statistic for other programs, provided they all assess impact on fluency and alphabetics.

An alternative solution is to judge the relative value of programs according to impact on the ultimate or distal goal of any reading intervention - to improve reading comprehension - even when programs target more proximal outcomes such as alphabetics. In our analysis such a comparison is not possible because only four of the seven programs measured impacts on reading comprehension and, of these four, three showed non-significant results. We recommend that future evaluations of early literacy programs include a common measure or measures of reading comprehension and, ideally, common measures of each of the literacy domains. Use of common measures would also eliminate the need to calculate effect sizes which, as previously mentioned, are hard to interpret and vulnerable to inflation when the population being studied is homogeneous.

For topic areas such as high school completion or college attendance where it is possible to monetize the outcomes of an intervention by predicting long term benefits including increased earnings and tax contributions, a cost-benefit analysis can solve the problem of valuing multiple outcomes simultaneously. While Hernandez (2012) demonstrated that third grade scores on a test of word recognition and pronunciation ability can predict high school graduation rates, such findings have not been used to conduct cost-benefit analyses of reading programs and we would not recommend this approach. In part, this is because it would assume the literacy domains were independent of each other but related to later long-term benefits. Additionally, there are other benefits to literacy that cannot be easily monetized, such as cultural

participation and enjoyment. Finally, there are stronger predictors of high school graduation than literacy, namely mathematics achievement (Duncan et al., 2007).

Bearing in mind the program differences, contextual details, and caveats we outline above, the cost-effectiveness ratios obtained for each literacy program we analyzed may be considered as one piece of information by education decision-makers when choosing among programs to implement in their schools. Other considerations, beyond the obvious one of actual costs, might include the relative importance of different educational goals, budget limits, needs of the student population, educational philosophy and capacity of the teaching body, union rules, parent preferences, and local availability of the personnel and other resources required to implement the program at the desired scale.

For future evaluations of early literacy studies, 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. We also suggest that evaluation standards require that programs be evaluated at multiple sites with large enough sample sizes to allow an investigation of whether and how effectiveness and costs vary across sites. Inclusion of qualitative descriptions of implementation to explain the challenges and advantages encountered at particular sites would also facilitate an assessment of whether and how resource use and resource management is 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.

Cost-effectiveness evaluations of early literacy programs that are widely used in American schools should be commissioned to allow decision-makers a realistic menu of options from which to choose when making resource allocation decisions. In general, across all educational program areas, government agencies and private foundations could put greater emphasis on the collection and use of cost data in combination with effectiveness data in making decisions regarding program funding.

REFERENCES

Abramson, P. (2011). The 2011 school construction report. School planning and management. Retrieved from: http://www.peterli.com/spm/pdfs/SchoolConstructionReport2011.pdf.

Allegretto, S., Corcoran, S. & Mishel, L. (2004). *How does teacher pay compare? Methodological challenges and answers*. Washington, D.C.: Economic Policy Institute.

Allegretto, S., Corcoran, S. & Mishel, L. (2008). *The teaching penalty. Teacher pay losing ground*. Washington, D.C.: Economic Policy Institute.

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

Bowden, A. B. (2013). *Cost-effectiveness analysis and the challenge of site-level variation: an evaluation of Talent Search*. Paper presented at the annual meeting of the Association for Education Finance and Policy, Washington, DC.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, 112(3), 579-620.

Chambers, J.G., Shkolnik, J. & Pérez, M. (2003). *Total expenditures for students with disabilities, 1999-2000: spending variation by disability*. Center for Special Education Finance. Retrieved from: http://www.csef-air.org/publications/seep/national/Final_SEEP_Report_5.PDF

Currie, J. (2001). Early childhood education programs. *The Journal of Economic Perspectives,* 15(2), 213-238.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *The American Economic Review*, 85(3), 341-364.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43(6), 1428-1446. doi: 10.1037/0012-1649.43.6.1428.

Fern, F. E., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, 23(2), 89-105.

Harris, D. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: combining effects with costs. *Educational Evaluation and Policy Analysis,* 31(1), 3-29.

Harris, D. & Sass, T. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8), 798-812.

Hernandez, D.J. (2012). *Double jeopardy: how third-grade reading skills and poverty influence high school graduation*. Annie E. Casey Foundation, Baltimore MD. Retrieved from: http://www.aecf.org/~/media/Pubs/Topics/Education/Other/DoubleJeopardyHowThirdGradeReadingSkillsandPovery/DoubleJeopardyReport030812forweb.pdf

Hill, C.J., Bloom, H.S., Black, A.R., & Lipsey, M.W. (2007). *Empirical benchmarks for interpreting effect sizes in research*. MDRC Working Paper.

Hummel-Rossi, B. & Ashdown, J. (2010). *Cost-Effectiveness analysis as a decision tool in selecting and implementing instructional interventions in literacy*. Reading Recovery Council of North America online document. Retrieved from: http://fdf.readingrecovery.org/component/content/article/154-cost-effectiveness-analysis

Iverson, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology*, 85(1), 112–126.

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, 95(2), 211-224.

Levin, H. M. (1975). Cost-effectiveness in evaluation research. In M.Guttentag and E. Struening eds. *Handbook of Educational Research* Vol.2, Chap. 5. Beverly Hills, CA: Sage Publications.

Levin, H. M. (1995). Cost-effectiveness analysis. In M.Carnoy (Ed.), *International encyclopedia of economics of education* (2nd ed. pp.381-386). Oxford, UK.

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

Levin, H. M. (2002). Issues in designing cost-effectiveness comparisons of whole school reforms. In H.M. Levin & P.J. McEwan (Eds.), *Cost-effectiveness and educational policy: 2002 yearbook of the American Education Finance Association* (pp. 71-96). New York: Eye on Education.

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 and London: Routledge.

Levin, H. M., Belfield, C. R., Hollands, F. M., Bowden, A. B., Cheng, H, Shand, R., Pan, Y., & Hanisch-Cerda, B. (2012). *Cost-effectiveness analysis of interventions that improve high school completion*. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

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

Metsala, J.L., & Walley, A.C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J.L. Metsala & L.C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Hillsdale, NJ: Erlbaum.

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.

National Center for Education Statistics. (2012). *The Condition of Education*. Retrieved from: http://nces.ed.gov/programs/coe/pdf/coe_tot.pdf

National Fire Protection Association (1999). *NFPA 101. Life safety code.* Retrieved from: http://www.nfpa.org.

Nelson, J. R., Benner, G. J., & Gonzalez, 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.

Pinnell, G. S., DeFord, D. E., & Lyons, C. A. (1988). *Reading Recovery: Early intervention for at-risk first graders (Educational Research Service Monograph)*. Arlington, VA: Educational Research Service.

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.

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

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1), 175-214.

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).

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21(4), 360-407.

Stein, M. L.; Berends, M.; Fuchs, D.; McMaster, K.; Sáenz, L.; Yen, L.; Fuchs, 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.

Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., Wexler, J. Francis, D. J, Rivera, M. O., Lesaux, N. (2007). *Academic literacy instruction for adolescents: A guidance document from the Center on Instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved online at http://www.centeroninstruction.org/files/Academic%20Literacy.pdf

Torgesen, J. K., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W.,Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2006). *Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers*. Volume II: National Assessment of Title I: Interim Report to Congress, Institute of Education Sciences. Retrieved from: http://www.ed.gov/rschstat/eval/disadv/title-iinterimreport/index.html.

Tsang, M.C. (1997). Cost analysis for improved educational policymaking and evaluation. *Educational Evaluation and Policy Analysis*, 19 (4), 318-324.

U.S. Department of Education, Office of Educational Research and Improvement (1997). *Time spent teaching core academic subjects in elementary schools: Comparisons across community, school, teacher and student characteristics*. Retrieved from: http://0-nces.ed.gov.opac.acc.msmc.edu/pubs/97293.pdf

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.

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2005). Relative effectiveness of reading practice or word-level instruction in supplemental tutoring: How text matters. *Journal of Learning Disabilities*, 38(4), 364–380.

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology*, 98(3), 508–528.

WWC Intervention Report: Fast ForWord. (2007a). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Intervention Report: Stepping Stones to Literacy. (2007b). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Intervention Report: Corrective Reading. (2007c). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Intervention Report: Wilson Reading System. (2007d). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Intervention Report: Reading Recovery. (2008). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Intervention Report: Sound Partners. (2010). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Evidence Review Protocol for Beginning Reading Interventions, Version 2.1. (2012a). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Intervention Report: Peer-assisted Learning/Literacy Strategies. (2012b). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

WWC Procedures and Standards Handbook, Version 3.0. (2013). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

APPENDIX I: DEFINITIONS OF LITERACY TERMS, DRAWING ON WWC DEFINITIONS

Alphabetic principle	The idea that written spellings systematically represent spoken words (Snow et al., 1998).
Alphabetics	The domain of alphabetics for beginning readers is composed of the constructs of phonemic awareness, phonological awareness, letter identification, print awareness, and phonics (WWC, 2012a). We also include word reading fluency, a construct that WWC does not specifically address, in this domain.
Letter knowledge	Knowledge of the names of the letters of the alphabet, also called letter identification (WWC, 2012a).[10]
Literacy construct	WWC categorizes each literacy domain (see definition below) into one or more constructs that are measured by different instruments. Constructs in the domain of alphabetics include phonemic awareness, phonological awareness, letter identification, print awareness, phonics. We also include word reading fluency as a construct within the domain of alphabetics. WWC constructs in the domain of comprehension include vocabulary development and reading comprehension, though we chose to separate these out given evidence that they are quite distinct (e.g., Snow et al., 1998). There is only one construct in the domains of reading fluency and general reading achievement, i.e., reading fluency and general reading achievement respectively (WWC, 2012a).
Literacy domain	WWC categorizes beginning reading skills into four domains, i.e., alphabetics (phonemic awareness, phonological awareness, letter identification, print awareness, and phonics),[11] reading fluency, comprehension (vocabulary development and reading comprehension), and general reading achievement (WWC, 2012a).
Phonemic awareness	"Phonemic awareness (or phoneme awareness) refers to the understanding that the sounds of spoken language—phonemes—work together to make words, and phonemes can be substituted and rearranged to create different words. Phonemic awareness includes the ability to identify, think about, and work with the individual sounds in spoken words. Phonemic awareness helps children learn how to read and spell by allowing them to combine or blend the separate sounds of a word to say the word (e.g., "/c/ /a/ /t/ – cat")" (WWC, 2012a, p.1).

10 WWC defines the construct "letter identification" in the evidence review protocol while using the wording "letter knowledge" in the intervention reports. We assume that the two concepts are interchangeable.
11 We also include word reading fluency, a construct that WWC does not address, in this domain.

Phonics	"Phonics refers to (a) the knowledge that there is a predictable relationship between phonemes (the sounds in spoken language) and graphemes (the letters used to represent the sounds in written language), (b) ability to associate letters and letter combinations with sounds and blend them into syllables and words, and (c) understanding that this information can be used to decode or read words. Spelling is included in the review as an acceptable phonics outcome" (WWC, 2012a, p.2).
Phonological awareness	"Phonological awareness is a more encompassing term than phoneme/phonemic awareness. It refers to phoneme awareness and to awareness of larger spoken units such as syllables and rhyming words. Tasks of phonological awareness might require students to generate words that rhyme, to segment sentences into words, to segment polysyllabic words into syllables, or to delete syllables from words (e.g., what is "cowboy" without "cow"?)" (WWC, 2012a, p.1).
Print awareness	"Print awareness refers to knowledge of concepts about print, such as (a) print carries a message; (b) print has conventions, such as directionality (left to right, top to bottom), differences between letters and words, distinctions between upper case and lower case, and punctuation; and (c) books have some common characteristics (e.g., author, title, front/back)" (WWC, 2012a, p.2).
Reading comprehension	"Reading comprehension refers to the understanding of the meaning of a passage. Reading comprehension depends on various underlying components, including decoding (the ability to translate text into speech), knowledge of word meanings, fluency (the ability to read text accurately and automatically), and the ability to understand and interpret spoken language. Struggling readers may have difficulty with any of these components of reading or with multiple components" (WWC, 2012a, p.2).
Reading fluency	"Fluency is the ability to read text accurately, automatically, and with expression (including appropriate pausing, response to punctuation, etc.) while extracting meaning from it" (WWC, 2012a, p.2). (We note that the measures used in the studies we reviewed assess the ability to read text accurately and automatically, but do not typically assess expression or limit these abilities to instances when children are extracting meaning from text).
Vocabulary	"Vocabulary development refers to the development of knowledge about the meanings, uses, and pronunciation of words. The development of receptive vocabulary (words understood) and expressive vocabulary (words used) is critical for reading comprehension" (WWC, 2012a, p.2).

APPENDIX II: LITERACY INTERVENTIONS INTERVIEW PROTOCOL (GENERIC)

Introduction to IES Cost-Effectiveness Study

The federal government's What Works Clearinghouse has identified Program X as an early literacy program that is effective in improving phonics.

Our research is funded by a federal grant from the Institute for Educational Sciences at the U.S. Department of Education. All responses will be anonymous and confidential as per IRB protocol #12-270.

Our main interest is to figure out what resources (or "ingredients") were used to implement the program at the sites included in the evaluation, e.g., amount of teacher/teaching aide time helping/supervising students, training time spent by teachers learning how to use the program, equipment and materials needed to deliver the program, how many students were served and for how much time.

Program name:

Site served: Interview date:

Start time: End time:

Interviewer name: Telephone number called:

Interviewee name: Current position:

Opening Questions

What is your current connection with [program X]?

Did you have any involvement in the program at this or any other site in 20XX (the time of the program evaluation)?

How many students were served at this site by program X in:

(Year 1) (Year 2), etc.

Please provide a general description of the activities the program entailed at your site at the time of the evaluation:

How many students were served by each activity? Did the activity serve the entire class, or some sub-set of students? Were students pulled out of class? If so, what was the rest of the class doing during this time and did the pulled-out students make up missed classroom time?

Over how many weeks did the program occur? How often did each activity occur? For how long each time? When? Was it all during the school day, or at least in part during times outside of regular school hours, like nights, weekends, or summers?

Where did each activity occur? Specifically, were the students in their regular classrooms, or some other location?

Who was directly involved in each activity? Were any personnel required above and beyond what would be required for ordinary classroom instruction?

Personnel

The questions listed below are intended to gather detailed data on personnel. The questions are organized by headings to indicate the level of administration: state, district, and school personnel (principals, teachers, counselors, specialists, aides, volunteers, coaches, tutors, etc.).

We are interested in any personnel involved in the program – planning, implementing (i.e. actually delivering the services), supervising, volunteering.

State:
Who was involved in the program at the state level? Who spent a substantial amount of his/her time running this program? (Both elected/appointed officials and civil servants).

District:
Who was involved in the program at the district level?

Role in the program?

Number of people in that position?

Time spent on the program? (hours per week, or percentage of time).

What background requirements (qualifications, years of experience) were required for the position?

Was there anyone else who was part of the program at the district level?

At the district level, what office space and equipment was provided for people who worked on the program?

What percentage of the time was that space used by the program?

How many square feet did the space occupy?

What training did the individuals in these positions receive? Did training occur in multiple years?

Did the individuals in these positions provide training to others?

School:

Administration

What was the principal's role in the program? (meetings, professional development, scheduling, etc.).

How much time did the principal spend on the program per year?

What are the principal's qualifications? (e.g., degree, years of experience).

How long has the principal (or other administrators involved in the program) been at the school? Does tenure at the school impact the principal's ability to administer the program effectively?

Teachers

What was a teacher's role in the program? (meetings, professional development, implementation, data collection, etc.).

How much time did the teacher spend on the program per week?

Was the time spent on the program evenly spread across all weeks of the implementation?

If not, how was it spread out?

Did the program require any teacher time outside of regular school hours? How does the allocation of teacher time under the program compare to what happens normally?

How were teachers selected to participate in the program?

Can you think of any special qualifications or characteristics about the teachers selected to implement the program? Did the program require teachers to have a particular level of experience to successfully implement it?

How long have the teachers been at the school? Does tenure at the school impact the teachers' ability to administer the program effectively?

Volunteers

Did the program involve any volunteers?

If so, how many volunteers? How often did they volunteer and for how long?

Did the volunteers need any special training, qualifications, or experience?

Other School-level Personnel

What other types of school-level employees worked on the program?

How many of these employees were working for the program per year?

Were these full time positions? lf not, explain:

Were there any requirements for obtaining these positions? (degree, experience, etc.).

What was the ratio of these employees to teachers? To students?

Did these employees provide services to the district as well?

Training

Was any training provided by Program X to any of the personnel in year 1 of the program implementation (give year)?

Who did the training?

What were the trainers' qualifications?

How were personnel trained in the first year?

Duration and frequency of training, e.g., number of days per year?

Timing – was it during the school day or after-school/summer training?

What facilities were used and where?

If training took place off site, were any of the following provided and if so for how long: lodging, travel, per diem?

What materials and equipment were used?

What training/professional development was provided beyond the first year? (repeat questions listed above for each year)

How many days per year were substitute teachers hired to cover teachers' classes because the teachers were involved in the program?

How many substitute teachers were hired?

Materials and equipment

What materials, such as student workbooks, teacher manuals, and school or office supplies, were required for the program? *[Note: The questions in this section can be made much more specific and detailed once an initial ingredients list is created for each program.]*

How many of each was required?

How often did they need to be replaced – for each item, was it something that you needed to purchase once, or did it need to be replenished on a regular basis?

Did the program require the use of computers/printers/projectors etc.?

If so, how many of each was required, for how long, and how often?

Were the computers in any sort of specialized location, like a lab, or were they located in the classroom?

Was internet access needed?

Did the program receive any contributed donations of materials, supplies, or equipment? If so, what donated materials were used by the program?

Facilities

(Try to identify square footage of each space)

Did the program take place in any location besides the regular classroom, or did it require the use of the classroom beyond the time ordinarily used for class?

If so, what spaces were required – for instance, a smaller classroom or office for small-group instruction or tutoring? How often were these spaces used, and how large were they?

Did the program require use of any office space, storage space, or meeting space at the school for administration or training purposes? If so, what spaces, how large were they, how often were they used and for how long?

What office space and equipment was provided for people who work on the program at the school?

What percentage of the time was that space used by the program?

Other questions

Did the program involve any travel for personnel or students? If so, who traveled, to where, using what mode of transportation, how often?

Did the program provide any additional goods or services to student participants, for example, as rewards or incentives for performance? This may include small prizes, food, field trips, movie tickets, etc.

Did the program require any inputs from students' families? For example, did parents need to come in for additional conferences?

Are there any other aspects of the program – including resources paid for by the school and other donated goods and services – that we haven't covered?

Do you know how differently your school implemented the program as compared to other schools, or compared to how the program was originally designed?

Is there anyone else you think we should contact at the state, district or school level who might be able to give us further information about the ingredients and costs of implementing Program X?

Do you know how much, if at all, the program may have changed since the evaluation in 20XX?

APPENDIX III: PRICES OF INGREDIENTS

The prices of ingredients were collected from databases, documents, and Web sites across a range of sources. From these sources we compiled a database of ingredients prices.

Personnel Pricing

The prices of personnel are the most important item in ingredients pricing. We reviewed many potential sources for data on personnel prices, including:

CPS (Current Population Survey, US Department of Labor). Reports median gross wages per occupation reported by occupied households. Reported wage is on an annual basis for every occupation estimated from an hourly wage considering 2,080 hours a year, except for teachers. For teachers the wage reported is annual. Teachers are asked to report their annual wage regardless the amount of hours they work. Teacher wages for year 2010 are available (see Allegretto, Corcoran, & Mishel, 2004). [http://www.bls.gov/bls/empsitquickguide.htm]

CPS-ORG (Current Population Survey – Outgoing Rotation Groups, US Department of Labor). Includes earnings data per occupation per educational level (see Allegretto, Corcoran, & Mishel, 2008). [http://www.nber.org/data/morg.html]

NCS (National Compensation Survey, US Department of Labor). Reports mean annual and weekly wages from survey of employers. NCS also reports employee benefits per hour worked for state/local government employees and civilian workers. Teacher wages and benefits for year 2010 are available but weekly wage estimates are sensitive to assumptions about weeks worked per year (see Allegretto et al., 2004). [http://www.bls.gov/ncs/]

CES (Current Employment Statistics, US Department of Labor). Reports industry – but not occupational – data on employment, hours, and earnings of workers on nonfarm payrolls. [http://www.bls.gov/ces/home.htm]

NEA Education Worker Survey (National Education Association). Reports mean wages per occupation for teachers and educational paraprofessionals. Data are collected from state departments of education. [http://www.nea.org/home/13566.htm]

SASS (Schools and Staffing Survey, US Department of Education, National Center for Education Statistics). The Public Teacher Questionnaire from SASS reports national data on teachers' wages by highest degree earned and years of experience. Data for primary and secondary school teachers is aggregated; results for some teacher qualifications and years of experience not reported due to low sample size. Data from Public and Private School Principal Questionnaire from SASS reports information on wages for K-12 principals by years of experience and institutional type. Data from 2007-2008. [http://nces.ed.gov/programs/digest/d11/tables/dt11_079.asp]

IPEDS (Integrated Postsecondary Education Dataset, US Department of Education, National Center for Education Statistics). Reports salaries for faculty in public and private institutions for the academic year on 11/12-month and 9/10-month contracts. No data is reported for salaries and benefits of other staff. [http://nces.ed.gov/ipeds/].

CUPA-HR (College and University Professional Association for Human Resources). Reports data on median salaries from annual survey for senior and mid-level college administrators in different types of institutions based on the 2005 Carnegie Classification. [http://www.cupahr.org/surveys]

OECD-INES (Indicators of Education Systems) Survey on Teachers and Curriculum (Education at a Glance 2011). Reports data on statutory teachers' salaries for 2008-2009. Statutory salary for a full-time teacher is the number of hours per year that a teacher is required to spend teaching. It does not adjust salaries for the amount of time that teachers spend on other teaching-related activities. [http://www.oecd-ilibrary.org/]

OES (Occupational Employment Statistics Survey, US Department of Labor). Reports mean annual wages reported by employers (but not disaggregated by educational levels or years of experience) and information on minimum levels of educational level required across occupations. Teacher wages for year 2010 are available. [http://www.bls.gov/oes/home.htm]

OOH (Occupational Outlook Handbook, Bureau of Labor Statistics, US Department of Labor). Reports median salaries for over 340 occupations that cover 85 percent of jobs in the economy. [http://www.bls.gov/ooh/home.htm]

NCS (National Compensation Survey, Bureau of Labor Statistics, US Department of Labor). Provides detailed information on average wages for over 800 occupations, measures of benefits, and employer cost for employee compensation. Three levels of data are reported in the survey: local, regional, and national. [http://www.bls.gov/ncs/home.htm]

Facilities Pricing

Facilities prices are important because educational buildings have specific requirements with respect to space, building codes, and provisions.

Data on prices for facilities is less detailed than for personnel and reports vary in terms of the type of costs they report and the categories of costs included. For example, costs may refer exclusively to construction costs and not include the costs of acquiring the site, site development, or furnishings and equipment. These prices are sensitive to location. For example, state policies on school construction differ because of state specification requirements (e.g., for earthquake-proof construction in California and for hurricane shelters in Florida). These prices are also sensitive to reporting year because the construction industry is sensitive to market trends. For instance, College Planning and Management magazine reports the 2011 price per square foot for technology buildings in postsecondary education at almost twice the price for 2010.

We reviewed several sources for information on national prices of school facilities. These included:

NCEF (National Clearinghouse for Educational Facilities). The NCEF recommends as sources of facility costs the Annual Construction Reports from the School Planning and Management and College and Planning and Management magazines. Their annual construction reports provide national and regional cost data on school and university construction. [http://www.ncef.org/ds/index.cfm]

School Planning and Management magazine reports total cost per construction reported. Construction costs are two-thirds of total costs per square foot (sq. ft.). Other costs include: site purchase (2%), site development (9%), furnishing and equipment (14%), fees/others

(8%). Also, square footage is not disaggregated by functionality. [http://www.peterli.com/spm/pdfs/SchoolConstructionReport2011.pdf]

College Planning and Management magazine reports construction costs and these are disaggregated in terms of functionality of the sq. ft. [http://www.peterli.com/cpm/pdfs/CollegeConstructionReport2011.pdf]

Reed Construction Data provides detailed information on building costs based mainly on cost estimation from 2008. Costs are separated for universities and community colleges, as well as for schools by level (elementary, middle, and high) and by use (classrooms, auditoriums, laboratories). [http://www.reedconstructiondata.com/building-types]

State statutes for California, New York, New Jersey, Washington, DC. State statutes include some information on the requirements for school facilities construction and/or instructional building aid.

Educational institutions vary in the square footage requirements. Per student in a school building, these are: 125 sq. ft. for elementary school; 149 sq. ft. per middle school; and 156 sq. ft. per high school (Abramson, 2011). Per student in a classroom, the occupancy load requires 20 sq. ft. (National Fire Protection Association, 1999).

The costs of facilities need to be amortized over their operational life. This lifespan may vary but is at least 25 years. According to the National Center for Education Statistics, in 1998 the average public school building in the United States was 42 years old. [http://nces.ed.gov/surveys/frss/publications/1999048/index.asp]

Materials and Equipment Pricing

The prices for materials and equipment vary according to type. For most materials, e.g. printers and computers, we derived prices from market prices using internet searches. We used retail prices net of transportation [http://www.educationmarketplace.com].

Most current prices were adjusted to 2010. In the particular case of technological equipment such as computers and laptops, prices were not adjusted as prices do not move in line with inflation. Materials and equipment were amortized using depreciation periods reported by the IRS [http://www.irs.gov/pub/irs-pdf/p946.pdf]

Database of Ingredients Prices

From the above sources we compiled a database of prices of ingredients. The database is available at www.cbcse.org under the Cost Resources tab. This database encompasses many items for each category of ingredient:

Personnel (190 items). These items vary with level (K-12, postsecondary), occupation, education, annual/weekly/hourly wage, full-time/part-time status, years of experience, mean/median salary, date of estimate, and data source.

Facilities (15 items). These items vary across level of institution, purpose of building, date of estimate, data source, and median sq. ft. for use versus new construction. They also vary with assumptions about amortization.

Materials/equipment (9 items). These items varied with the type of equipment, and duration of rental.

APPENDIX IV: ABBREVIATIONS

FFW1	Fast ForWord Reading 1
K-PALS	Kindergarten Peer-Assisted Learning Strategies
na	not applicable (used by WWC when an average effect size is reported which may include mixed results).
nm	not measured
ns	not significant
WWC	What Works Clearinghouse