Success for All: A Quantitative Synthesis of U. S. Evaluations

Alan C. K. Cheung

Department of Educational Administration and Policy

Faculty of Education

The Chinese University of Hong Kong

Chen Xie

Institute of Curriculum and Instruction

Faculty of Education

East China Normal University

Tengteng Zhuang

Department of Educational Administration and Policy

Faculty of Education

The Chinese University of Hong Kong

Amanda J. Neitzel

Johns Hopkins University

Robert E. Slavin

Johns Hopkins University

Abstract

Success for All (SFA) is a comprehensive whole-school approach designed to help high-

poverty elementary schools increase the reading success of their students. It is designed to

ensure success in grades K-2 and then build on this success in later grades. SFA combines

instruction emphasizing phonics and cooperative learning, one-to-small group tutoring for

students who need it in the primary grades, frequent assessment and regrouping, parent

involvement, distributed leadership, and extensive training and coaching. Over a 33-year

period, SFA has been extensively evaluated, mostly by researchers unconnected to the

program. This quantitative synthesis reviews the findings of these evaluations. Seventeen

U.S. studies meeting rigorous inclusion standards had a mean effect size of +0.24 ($p < .05$) on

independent measures. Effects were largest for low achievers (ES = +0.54, $p < .01$). Although

outcomes vary across studies, mean impacts support the effectiveness of Success for All for

the reading success of disadvantaged students.

Success for All: A Quantitative Synthesis of Evaluations

The reading performance of students in the United States is a source of deep concern. American students perform at levels below those of many peer nations on the Program for International Student Assessment (PISA; OECD, 2019). Most importantly, there are substantial gaps in reading skills between advantaged and disadvantaged students, between different ethnic groups, and between proficient speakers of English and English learners (NCES, 2019). These gaps lead to serious inequalities in the American economy and society. America's reading problem is far from uniform. On PISA Reading Literacy (OECD, 2019), American 15 year old students in schools with fewer than 50% of students qualifying for free lunch scored higher than those in any country. The problem in the U.S. is substantially advancing the reading skills of students in high-poverty schools. The students in these schools are capable of learning at high levels, but they need greater opportunities and support to fully realize their potential.

Research is clear that students who start off with poor reading skills are unlikely to recover without significant assistance (Cunningham & Stanovich, 1997; National Reading Panel, 2000). A study by Lesnick, George, Smithgall, & Gwynne (2010) found that students reading below grade level in third grade were four times more likely than other students to drop out before high school graduation.

Evidence about the role of early reading failure in long-term school failure (e.g., National Reading Panel, 2000) has led to a great deal of research and development focused on ensuring that students succeed in reading in the elementary grades. Recent reviews of programs for struggling readers by Neitzel, Lake, Pellegrini, & Slavin (2020b) and Wanzek et al. (2016) have identified many effective approaches, especially tutoring and professional development strategies. However, in high-poverty schools in which there may be many students at risk of reading failure, a collection of individual approaches may be insufficient or

inefficient. In such schools, whole school, coordinated approaches may be needed to ensure that all students succeed in reading.

Success for All

Success for All (SFA) was designed and first implemented in 1987 in an attempt to serve very disadvantaged schools, in which it is not practically possible to serve all struggling readers one at a time. The program emerged from research at Johns Hopkins University, and since 1996 has been developed and disseminated by a non-profit organization, the Success for All Foundation (SFAF). SFA was designed from the outset to provide research-proven instruction, curriculum, and school organization to schools serving many disadvantaged students.

Theory of Action. Success for All was initially designed in a collaboration between researchers at Johns Hopkins University (JHU) and leaders of the Baltimore City Public Schools (BCPS), whose high-poverty schools had large numbers of students falling behind in reading in the early elementary grades, losing motivation, and developing low expectations for themselves. Ultimately, these students entered middle school lacking basic skills and, in too many cases, no longer believing that success was possible. The JHU-BCPS team was charged with developing a whole-school model capable of ensuring success from the beginning of students' time in school. The theory of action the team developed focused first on ensuring that students were successful in reading in first grade, providing a curriculum with a strong emphasis on phonemic awareness and phonics (National Reading Panel, 2000; Shaywitz & Shaywitz, 2020; Snow et al., 1998), and using proven instructional methods such as cooperative learning (Slavin, 2017), and effective classroom management methods (e.g., Good & Brophy, 2018). Students in grades 1-5 are grouped by reading level across grade lines, so that all reading teachers had one reading group. For example, a reading group at the 3-1 level (third grade, first semester) might contain some high-performing second graders,

many third graders, and some low-performing fourth graders, all reading at the 3-1 level. Students in the primary grades, but particularly first graders, may receive daily, 30-minute computer-assisted tutoring, usually in groups of four, to enable most struggling readers to keep up (Neitzel, Lake, Pellegrini, & Slavin, 2020b; Wanzek et al., 2016).

The core focus of the SFA model is to make certain that every student succeeds in basic reading. In addition to the reading instruction and tutoring elements, students who need them can receive services to help them with attendance, social-emotional development, parent involvement, and other needs. After students reach the 2-1 reading level, they continue to receive all program services except tutoring. The upper-elementary program is an adaptation of Cooperative Integrated Reading and Composition (CIRC; Stevens, Madden, Slavin, & Farnish, 1987). The design of the SFA program in at reading levels 3-5 is focused on cooperative learning, comprehension, metacognitive skills, and writing.

The theory of action for SFA, therefore, assumes that students must start with success, whatever this takes, in the expectation that early success builds a solid base for later learning, positive expectations for future success, and motivation to achieve. However, success in the early grades is seen as necessary, but not sufficient. Evidence on the difficulties of ensuring long-term maintenance of reading gains from highly successful first grade tutoring programs (e.g., Blachman et al., 2014; Hurry & Sylva, 2007) demonstrate that ensuring early-grade success in reading cannot be assumed to ensure lifelong reading success. The designers of SFA intended to build maintenance of first-grade effects by continuing high-quality instruction and classroom organization after an intensive early primary experience sets students up for success. Beyond reading and tutoring, the design seeks to build on students' strengths by involving their parents, teaching social-emotional skills, and ensuring high attendance.

Figure 1 summarizes the SFA theory of action. At the center is success in reading in

grades K-2, and then 3-5. All other components of the model support these outcomes. Only tutoring is limited to Grades 1-2. Other elements continue through the grades.

The logic of Success for All is much like that of response to intervention (Fuchs & Fuchs, 2006), now often called Multi-Tier Systems of Support (MTSS). That is, teachers receive extensive professional development and in-class coaching to help them use proven approaches to instruction and curriculum. Students who do not succeed despite enhanced teaching may receive one-to-small group or, if necessary, one-to-one tutoring. Ongoing assessment, recordkeeping, and flexible grouping are designed to ensure that students receive instruction and supportive services at their current instructional level, as they advance toward higher levels. Program components focus on parent involvement, classroom management, attendance, and social-emotional learning, to solve problems that may interfere with students' reading and broader school success. Each school has a full-time facilitator to help manage professional development and other program elements, some number of paraprofessional tutors, and coaches from the non-profit Success for All Foundation, who visit schools approximately once a month to review the quality of implementation, review data, and introduce additional components.

Program Components

Success for All is a whole-school model that addresses instruction, particularly in reading, as well as schoolwide issues related to leadership, attendance, school climate, behavior management, parent involvement, and health (see Slavin, Madden, Chambers, & Haxby, 2009, for more detail). The program provides specific teacher and student materials and professional development to facilitate use of proven practices in each program component.

Literacy instruction. Learning to read and write effectively is essential for success in school. Success for All provides in-depth support for reading acquisition. Instructional

practices, teacher's guides, student materials, assessments, and job-embedded professional development are combined to create a comprehensive reading program.

The Success for All reading program is based on research and effective practices in beginning reading (e.g., National Reading Panel, 2000), and appropriate use of cooperative learning to enhance motivation, engagement, and opportunities for cognitive rehearsal (Slavin, 2017; Stevens, Madden, Slavin, and Farnish, 1987).

Regrouping. As noted earlier, students in grades one and up are regrouped for reading. The students are assigned to heterogeneous, age-grouped classes most of the day, but during a regular 90-minute reading period they are regrouped by reading performance levels into reading classes of students all at the same level. For example, a reading class taught at the 2-1 level might contain first, second, and third grade students all reading at the same level. The reading classes are smaller than homerooms because tutors and other certified staff (such as librarians or art teachers) teach reading during this common reading period.

Regrouping allows teachers to teach the whole reading class without having to break the class into reading groups. This greatly reduces the time spent on seatwork and increases direct instruction time. The regrouping is a form of the Joplin Plan, which has been found to increase reading achievement in the elementary grades (Slavin, 1987).

Preschool and kindergarten. Most Success for All schools provide a half-day preschool and/or a full-day kindergarten for eligible students. Research supports a balance between development of language, school skills, and social skills (Chambers, Cheung, & Slavin, 2016). The SFA preschool and kindergarten programs provide students with specific materials and instruction to give them a balanced and developmentally appropriate learning experience. The curriculum emphasizes the development and use of language. It provides a balance of academic readiness and non-academic music, art, and movement activities in a series of thematic units. Readiness activities include use of language development activities and Story

Telling and Retelling (STaR), which focuses on the development of concepts about print as well as vocabulary and background knowledge. Structured phonemic awareness activities prepare students for success in early reading. Big books as well as oral and written composing activities allow students to develop concepts of print story structure. Specific oral language experiences are used to further develop receptive and expressive language.

Curiosity Corner, Success for All's pre-kindergarten program, offers theme-based units designed to support a language-rich half-day program for three- and four-year olds that supports the development of social emotional skills and early literacy.

KinderCorner offers a full-day theme-based kindergarten program designed to support the development of oral language and vocabulary, early literacy, and social and emotional skills needed for long term success. KinderCorner provides students with materials and instruction designed to get them talking using cooperative discussion with an integrated set of activities. Opportunities for imaginative play increase both self-regulation and language. Formal reading instruction is phased in during kindergarten. Media-based phonemic awareness and early phonics ease students into reading, and simple but engaging phonetically regular texts are used to provide successful application of word synthesis skills in the context of connected text.

Beginning reading. Reading Roots is a beginning reading program for grades K-1. It has a strong focus on phonemic awareness, phonics, and comprehension (Shaywitz & Shaywitz, 2020; Snow et al., 1998). It uses as its base a series of phonetically regular but interesting minibooks and emphasizes repeated oral reading to partners as well as to the teacher. The minibooks begin with a set of "shared stories," in which part of a story is written in small type (read by the teacher) and part is written in large type (read by the students). The student portion uses a phonetically controlled vocabulary. Taken together, the teacher and student portions create interesting, worthwhile stories. Over time, the teacher portion

diminishes and the student portion lengthens, until students are reading the entire book. This scaffolding allows students to read interesting stories when they only know a few letter sounds.

Letters and letter sounds are introduced in an active, engaging set of activities that begins with oral language and moves into written symbols. Individual sounds are integrated into a context of words, sentences, and stories. Instruction is provided in story structure, specific comprehension skills, metacognitive strategies for self-assessment and self-correction, and integration of reading and writing. Brief video segments use animations to reinforce letter sounds, puppet skits to model sound blending, and live action skits to introduce key vocabulary.

Adaptations for Spanish speakers. Spanish bilingual programs use an adaptation of Reading Roots called Lee Conmigo ("Read With Me"). Lee Conmigo uses the same instructional strategies as Reading Roots, but is built around shared stories written in Spanish. SFA also has a Spanish-language kindergarten program, Descubre Conmigo ("Discover with Me"). Students who receive Lee Conmigo typically transition to the English SFA program in Grades 2 or 3, using special materials designed to facilitate transition. Schools teaching English learners only in English are provided with professional development focused on supporting the language and reading development of English learners.

Upper elementary reading. When students reach the second grade reading level, they use a program called Reading Wings, an adaptation of Cooperative Integrated Reading and Composition (CIRC) (Stevens, Madden, Slavin, & Farnish, 1987). Reading Wings uses cooperative learning activities built around story structure, prediction, summarization, vocabulary building, decoding practice, and story-related writing. Students engage in partner reading and structured discussion of stories or novels, and work toward mastery of the vocabulary and content of the story in teams. Story-related writing is also shared within

teams. Cooperative learning both increases students' motivation and engages students in cognitive activities known to contribute to reading comprehension, such as elaboration, summarization, and rephrasing (see Slavin, 2017). Research on CIRC has found it to significantly increase students' reading comprehension and language skills (Stevens et al., 1987).

Reading tutors. A critical element of the Success for All model is the use of tutoring, the most effective intervention known for struggling readers (Neitzel, Lake, Pellegrini, & Slavin, 2020b; Wanzek et al., 2016). In the current version of SFA, computer-assisted tutoring is provided by well-qualified paraprofessionals to groups of four children with reading problems. However, students with very serious problems may receive one to two or one to one tutoring. The tutoring occurs in 30-minute sessions during times other than reading or math periods.

Leading for success. Schools must have systems that enable them to assess needs, set goals for improvement, make detailed plans to implement effective strategies, and monitor progress on a child by child basis. In Success for All, the tool that guides this schoolwide collaboration is called Leading for Success.

Leading for Success is built around a distributed leadership model, and engages all school staff in a network of teams that address key areas targeted for continuous improvement. The leadership team manages the Leading for Success process and convenes the staff at the beginning of the school year and at the end of each quarter to assess progress and set goals and agendas for next steps. Staff members participate in different teams to address areas of focus that involve schoolwide supports for students and families as well as support for improving implementation of instructional strategies to increase success.

Schoolwide solutions teams. A Parent and Family Involvement Team works towards good relations with parents and to increase involvement in the school. Team members

organize "welcome" visits for new families, opportunities for informal chats among parents and school staff members, workshops for parents on supporting achievement and general parenting issues, and volunteer opportunities. Solutions teams also focus on improving attendance and intervening with students having learning and behavioral problems.

Program facilitator. A program facilitator works at each school to oversee (with the principal) the operation of the Success for All model. The facilitator helps plan the Success for All program, helps the principal with scheduling, and visits classes and tutoring sessions frequently to help teachers and tutors with individual problems. He or she works directly with the teachers on implementation of the curriculum, classroom management, and other issues, helps teachers and tutors deal with any behavior problems or other special problems, and coordinates the activities of the Family Support Team with those of the instructional staff.

Teachers and teacher training. Professional development in Success for All emphasizes on-site coaching after initial training. Teachers and tutors receive detailed teacher's manuals supplemented by three days of in-service at the beginning of the school year, followed by classroom observations and coaching throughout the year. For classroom teachers of grades one and above and for reading tutors, training sessions focus on implementation of the reading program (either Reading Roots or Reading Wings), and their detailed teacher's manuals cover general teaching strategies as well as specific lessons. Preschool and kindergarten teachers and aides are trained in strategies appropriate to their students' preschool and kindergarten models. Tutors later receive two additional days of training on tutoring strategies and reading assessment.

Throughout the year, in-class coaching and in-service presentations focus on such topics as classroom management, instructional pace, and cooperative learning. Online coaching is also used after coaches and teachers have built good relationships.

Special education. Every effort is made to deal with students' learning problems

within the context of the regular classroom, as supplemented by tutors. Tutors evaluate students' strengths and weaknesses and develop strategies to teach in the most effective way. In some schools, special education teachers work as tutors and reading teachers with students identified as learning disabled, as well as other students experiencing learning problems who are at risk for special education placement. One major goal of Success for All is to keep students with learning problems out of special education if at all possible, and to serve any students who do qualify for special education in a way that does not disrupt their regular classroom experience (see Borman & Hewes, 2002).

Consistency and variation in implementation. Success for All is designed to provide a consistent set of elements to each school that selects it. On engaging with schools, school and district staff are asked to agree to implement a set of program elements that the developers have found to be most important. These include the following:

- A full-time facilitator employed by the school. Typically, the facilitator is an experienced teacher, Title I master teacher, or vice principal already on the school staff, whose roles and responsibilities are revised to focus on within-school management of the SFA process.

- At least one full-time tutor, usually a teaching assistant, to work primarily with first graders who are struggling in reading.

- Implementation of the SFA KinderCorner (or Descubre Conmigo) program in kindergarten, Reading Roots (or Lee Conmigo) in grades one and two, and Reading Wings in grades 2 and above (for students who have tested out of Reading Roots). KinderCorner and Reading Roots are complete early reading approaches, but Reading Wings is built around widely used traditional or digital texts and/or trade books selected by schools.

- Professional development by SFA coaches, consisting of two days for all teachers, plus monthly on-site visits by SFA coaches.

•Regrouping for reading. During a daily 90-minute reading period, students are regrouped for reading starting in grade 1, as described above.

These elements are considered essential to SFA, and SFAF does not engage with schools that decline to implement and maintain all of them. After program inception, it of course occurs that schools cannot keep to their initial commitments, and some accommodations have to be made. For example, a school under financial pressure may have to use a half-time facilitator rather than full-time.

With respect to other elements of SFA, such as leadership, parent involvement, and special education policies, SFAF negotiates variations to accommodate school characteristics and district policies. As a result of its strong emphasis on consistency, the program elements believed to be most essential to reading outcomes do not vary significantly from school to school.

Evolution of program components over time. The basic design and operation of Success for All has remained constant for its entire 33-year history, but there has been constant change in the specific components. These are introduced because of learnings from experiences in schools, demand from schools and districts, findings of research, external grants, and advances in technology (see Peurach, 2011). The Reading Roots (K-1) reading program, for example, developed technology to help teachers present lessons and manage regular assessments. Reading Wings (2-5) has increasingly focused on the teaching of reading comprehension using metacognitive strategies. The tutoring program has evolved substantially. The main driver has been a quest for cost-effectiveness, as tutoring is expensive. Initially, tutoring was done by certified teachers one-to-one. However, this was not economically sustainable for most schools, so in the mid-1990s, SFAF developed a new model appropriate for use by teaching assistants. In the 2000s, SFAF began to introduce computer-assisted tutoring, taking advantage of increasing availability of computers in

schools. SFAF then began to develop and evaluate small group tutoring. In 2016, SFAF developed a computer-assisted small-group model that teaching assistants could use reliably with success in groups of four. This model requires one-eighth the personnel costs per tutored student of our original model, and gets equal outcomes, so it allows schools to serve many more students for the same cost (Madden & Slavin, 2017).

Some whole programs have been added, to enable SFAF to serve additional populations. SFAF added a preschool program in the mid-1990s, and added Spanish bilingual and sheltered English program around the same time. SFAF added the Leading for Success component in the 2000s, to improve schools' capacities to distribute leadership among its staff.

Any program as comprehensive as Success for All has to evolve to keep up with the times and to constantly improve its outcomes and reduce its cost and complexity. Success for All has always learned from its partners and its own staff, and incorporates these learnings continuously, in ways large and small.

Research on Success for All

Success for All has been in existence for 33 years, and currently (2020) provides services to about 1000 schools in the U.S. About half of these use the full program, and half use major components (most often, the K-2 reading program). The program has placed a strong emphasis on research and evaluation, and has always carried out or encouraged experimental or quasi-experimental evaluations to learn how the program is working and what results it is achieving for which types of students and settings. Studies of Success for All have usually been done by third party evaluators (i.e., researchers unrelated to the program developers). They have taken place in high-poverty schools and districts throughout the United States.

The present synthesis of research on Success for All includes every study of reading

outcomes carried out in U.S. schools that evaluated the program using methods that meet a set of inclusion standards described below. The purpose of the synthesis is to summarize the evidence and to identify moderators of program effects, and then to consider the implications of the findings for theory, practice, and policy.

The Need for a Meta-Analysis on SFA

Over the past 33 years, SFA programs have been widely applied and evaluated throughout the United States to help youngsters with their reading progress. However, these reports only focus on single evaluations of the intervention rather than synthesizing studies of all high-quality experiments over time. A meta-analysis of SFA studies was reported as part of a meta-analysis of comprehensive school reforms by Borman et al. (2003), and another meta-analysis was part of a synthesis of research on elementary reading programs by Slavin, Lake, Chambers, Cheung, & Davis (2009). SFA outcomes for struggling readers were included in a synthesis on that topic by Neitzel, Lake, Pellegrini, & Slavin (2020b). However, the present synthesis is the first to focus in detail on Success for All alone, enabling much more of a focus on its evidence base than was possible in reviews of many programs. Also, the review uses up-to-date methods for quantitative synthesis (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Piggott & Polanin, 2020).

The main objective of the current meta-analysis is to investigate the average impact of SFA on reading achievement. The three key main research questions are as follows:

1. What is the overall effect of SFA on student reading achievement?

2. Are there differential impacts of SFA on the reading achievement of different subgroups of students?

3. What study features moderate the effects of SFA on reading achievement?

Methods

Data Sources

The document retrieval process consisted of several steps (see Figure 2). The research team employed various strategies to identify all possible studies that have been done to evaluate reading outcomes of SFA. First, the team carried out a broad literature search. Electronic searches were made of educational databases (ERIC, Psych INFO, Dissertation Abstracts) using different combinations of key words (for example, "Success for All", "SFA", "reading", "Comprehensive School Reform") and the years 1989-2020. In addition, previous meta-analyses on reading interventions were searched and the reference lists of these meta-analyses were examined to identify any SFA studies. The authors contacted the Success for All Foundation, the developer of the program, to identify studies that might have been missed in the search, especially unpublished studies. Articles from any published or unpublished source that met the inclusion standards were independently read and examined by at least two researchers. Any disagreements in coding were resolved by discussion, and additional researchers read any articles on which there remained disagreements.

Inclusion and Exclusion Criteria

Criteria for inclusion and exclusion of studies were similar to those of the What Works Clearinghouse (WWC, 2020). They are as follows.

1. The studies evaluated SFA programs used in elementary schools. Studies had to appear between 1989 and 2020.

2. Studies had to be of students who started SFA in grades pre-K, K, or 1, as most tutoring (a key element of the theory of action) takes place in first grade.

3. The studies compared children taught in schools using SFA with those in control schools using an alternative program or standard methods.

4. Random assignment or matching with appropriate adjustments for any pretest differences (e.g., analyses of covariance) had to be used. In randomized experiments, a number of schools volunteered to participate, and half were assigned at random to use SFA,

while the remaining schools continued using existing methods. In matched studies, schools assigned to use SFA were matched in advance with control schools on factors such as pretests, poverty indicators, ethnicity, and school size. Post-hoc studies in which matching was done after experimental and control schools completed implementation were excluded. Studies without control groups, such as pre-post comparisons and comparisons to "expected" scores, were also excluded.

5. Pretest data had to be provided. Studies with pretest differences of more than 25% of a standard deviation were excluded, as required by WWC (2020) standards.

6. The dependent measures included quantitative measures of reading performance not created by SFA developers or researchers.

7. A minimum study duration of one school year was required.

8. Studies had to have at least two schools in each treatment group. This criterion avoided having treatment and school effects be completely confounded.

9. Study reported results at the end of the intervention period (for the main analyses) or interim results (for exploratory analyses examining impacts over time).

Coding

Studies that met the inclusion criteria were coded by one of the study team members and verified by another study team member. The fully coded data are available on GitHub (Neitzel et al. 2020a). Data to be coded beyond outcome measures, sample sizes, and effect sizes included substantive factors, methodological factors, and extrinsic factors. These are described below.

Substantive factors. Substantive factors describe the intervention, population, and context of the study.   These coded factors included duration of intervention, student grade level, and population description (race, ethnicity, English learner status, and free/reduced price meals status). Schools were categorized as being primarily African-American, Hispanic,

or White if more than half of students were of that race (or if there were subgroup analyses by race). They were considered high-poverty if at least 66% of students qualified for free lunch.

Methodological factors. Methodological factors included the research design (randomized or quasi-experimental design), and the type of outcome. Outcomes were categorized into three groups: general reading/comprehension, fluency, or alphabetics (WWC, 2014). Alphabetics includes subskills of reading such as letter identification and phonics outcomes, fluency includes reading accuracy and reading with expression, and comprehension outcomes assess the ability to understand connected text. General reading includes all types of reading outcomes. Comprehension is weighted heavily in general reading measures, so we combined general reading and comprehension scores into a single factor. The reading posttest scores used as the main outcome measures were those reported from the final year of implementation for a given cohort. For example, in a 3-year study with a K-2 and a 1-3 cohort, the third-year scores in grades 2 and 3 would be the main outcomes, and these would be averaged to get a study mean.

Extrinsic factors. Extrinsic factors coded included publication status, year of publication, and evaluator independence. Studies were considered independent if the list of authors did not include any of the original developers of SFA.

Statistical Analysis

The effect sizes of interest in this study are standardized mean differences. These are effect sizes that quantify the difference between the treatment and control group on outcome measures, adjusted for covariates, divided by standard deviations. This allows the magnitude of impacts to be compared across interventions and outcome measures. Effect sizes were calculated as the difference between adjusted posttest scores for treatment and control students, divided by the unadjusted standard deviation of the control group. Alternative procedures were used to estimate effect sizes when unadjusted posttests or unadjusted

standard deviations were not reported (Lipsey and Wilson, 2001).

In meta-analysis models, studies were weighted, to give more weight to studies with the greatest precision (Hedges, Tipton, & Johnson, 2010). In practice, this primarily involves weighting for sample size. Weights for each study were calculated according to the following formula:

$$W_j = \frac{1}{k_j(\bar{v}_j + \tau^2)}$$

where W_j is the weight for study j, k_j is the number of findings in study j, $\bar{v}_j$ is the average finding-level variance for study j, and τ^2 is the between-study variance in the study-average effect sizes (Hedges et al., 2010; Tipton, 2015). Variance estimates were adjusted for studies with cluster-level assignment, using the total variance for unequal cluster sample sizes (Hedges, 2007).

We used a multivariate meta-regression model with robust variance estimation (RVE) to conduct the meta-analysis (Hedges et al., 2010). This approach has several advantages. First, our data included multiple effect sizes per study, and robust variance estimation accounts for this dependence without requiring knowledge of the covariance structure (Hedges et al., 2010). Second, this approach allows for moderators to be added to the meta-regression model and calculates the statistical significance of each moderator in explaining variation in the effect sizes (Hedges et al., 2010). Tipton (2015) expanded this approach by adding a small-sample correction that prevents inflated Type I errors when the number of studies included in the meta-analysis is small or when the covariates are imbalanced. We estimated three meta-regression models. First, we estimated a null model to produce the average effect size without adjusting for any covariates. Second, we estimated a meta-regression model with the identified moderators of interest and covariates. Both the first and second models included only the outcomes at the end of the intervention period. Third, we

estimated an exploratory meta-regression model including the same identified moderators of interest and covariates, but that added results from interim reports, to better explore the change in impact over time. Both of the meta-regression models took the general form:

$$T_{ij} = \beta_0 + \beta_k X_{ij} + \beta_m X_j + \eta_j + \varphi_{ij} + \varepsilon_{ij}$$

where T_{ij} is the effect size estimate i in study j, β_0 is the grand mean effect size for all studies, β_k is a vector of regression coefficients for the covariates at the effect size level, X_{ij} is a vector of covariates at the effect size level, β_m is a vector of regression coefficients at the study level, and X_j is a vector of covariates at the study level, η_j is the study-specific random effect, and φ_{ij} is the effect size specific random effect. The X_{ij} and X_j included substantive, methodological, and extrinsic factors, as outlined above. All moderators and covariates were grand-mean centered to facilitate interpretation of the intercept. All reported mean effect sizes come from this meta-regression model, which adjusts for potential moderators and covariates. The packages metafor (Viechtbauer, 2010) and clubSandwich (Pustejovsky, 2020) were used to estimate all random-effects models with RVE in the R statistical software (R Core Team, 2020).

Results

Since first implemented in Baltimore in 1987, over 60 studies have been carried out to examine the effectiveness of SFA. However, only 17 studies met the inclusion criteria for this review. Common reasons for exclusion (see Online Appendix 1) included failure to have at least two schools in each treatment condition ($k = 17$), no appropriate data, or non-equivalent or missing pretests ($k = 13$), non-U.S. locations ($k = 17$), program started after first grade ($k = 2$), comparing to normed performance ($k = 2$), or comparing two forms of SFA ($k=4$).

Characteristics of Studies

The majority of the included studies were quasi-experiments (k = 15), and only two

were randomized studies. Three of the included studies were published articles and 14 were unpublished documents such as dissertations, conference papers, and technical reports. In terms of the relationship of the developer to the evaluator, most of the studies were determined to be independent ($k = 13$), while the remaining studies included at least one of the developers in the author list of the study ($k = 4$). All but one of the studies took place in schools with very high levels of economic disadvantage, with at least 66% of students receiving free or reduced-price lunches ($k = 16$).

Across these 17 studies, a total of 221 separate effect sizes were coded, with an average of 13 effect sizes per study. Six studies reported final effect sizes after 1 year (n = 55), 3 studies reported effect sizes after 2 years (n = 20), and 9 studies reported effect sizes after 3 or more years (n = 146). Six studies reported 85 outcomes for African-American students, either by reporting on a predominantly African-American student sample or by reporting on outcomes for African-American students separately, within a heterogeneous sample. Outcomes for Hispanic students were reported in 3 studies ($n = 34$). One study reported outcomes for White students ($n=4$). Four studies reported outcomes separately for English Learners (ELs), while eight studies reported on outcomes for low achievers separately. Outcomes were mainly of general reading or comprehension measures ($n = 90$) and alphabetics ($n = 97$), with fewer findings reported on fluency measures ($n = 34$).

Overall Effects

The results for the null model and full meta-regression model is shown in Table 1, which lists the two randomized studies and then all quasi-experiments in order of school sample size. The meta-regression model controlled for research design, independence of evaluator, duration of study, race/ethnicity of students, language status of students, baseline achievement level, and outcome type. There was an overall positive impact of SFA on reading achievement across all qualifying studies (ES = +0.24, p < .05). However, these outcomes vary

considerably, with a 95% prediction interval of -0.27 to +0.75. The prediction interval provides a sense of the heterogeneity of the outcomes, with 95% of the effect sizes in the population expected within this range. Study characteristics and findings of the 17 included studies are summarized in Table 2, and more detailed study-by-study information is shown in Appendix 2 in the online appendix.

Only two of the studies of SFA were large-scale cluster randomized experiments. Borman et al. (2007) carried out the first randomized, longitudinal study. Forty-one schools (21T, 20C) throughout the U.S. were randomly assigned to either the treatment or control condition. Children were pretested on the PPVT and then individually tested on the Woodcock Reading Mastery Test each spring for three years, kindergarten to second grade. At the end of this 3-year study, 35 schools and over 2,000 students remained. Using pretests as covariates, the HLM results indicated that the treatment schools significantly outperformed the controls on all three outcome measures, with an overall effect size of +0.25 ($p < 0.05$). The effect sizes were +0.22, +0.33, and +0.21 for Word Identification, Word Attack, and Passage Comprehension, respectively.

The second large-scale cluster randomized longitudinal study was carried out by Quint et al. (2015). Similar to the Borman study, 37 low-SES schools from five school districts in the U.S. were randomly assigned to treatment (N = 19) or control conditions (N = 18). Students were followed from kindergarten to second grade. The treatment schools scored significantly higher than the controls on phonics skills for second-graders who had been in the treatment group for all three years. No statistically significant differences were found on reading fluency and comprehension posttests.   However, among the lowest-performing students at pretest, those in the treatment group scored significantly higher than their counterparts in the control group on phonics skills, word recognition, and reading fluency.

All other U.S. studies of SFA used quasi-experimental designs, in which schools were

matched at pretest based on pretests and demographics, and then students in both groups were assessed each year, for from one to five years. Most of these quasi-experiments involved small numbers of schools, and would not have had sufficient numbers of clusters (schools) for adequate statistical power on their own. However, this meta-analysis combines these with other studies, weighting for sample size and other covariates, to obtain combined results that are adequately powered.

One of the QEDs was notable for its large size and longitudinal designs. Slavin et al. (1993; also see Madden et al., 1993) evaluated the first five schools to use Success for All. The schools, all high-poverty schools in Baltimore, were each matched with control schools with very similar pretests and demographics. All students were African American and virtually all students qualified for free lunches. Within schools, individual students were matched with control students. Students were followed from first grade onward, in a total of five cohorts. The mean effect size across all five cohorts was +0.59 for all students and +1.17 for low achievers. The mean effect size for fifth graders who had been in treatment or control schools since first grade was +0.46 overall and +1.01 for low achievers. A follow-up study of these schools was carried out by Borman & Hewes (2002). It obtained data from three cohorts of students followed to the eighth grade, so students would have been out of the K-5 SFA schools for at least three years. Results indicated lasting positive effects on standardized reading achievement measures (ES = +0.29), and SFA students were significantly less likely to have been retained in elementary school (ES = +0.39) or assigned to special education (ES=+0.18), in comparison to controls.

The second major, large-scale QED was a part of the University of Michigan's Study of Instructional Improvement (Rowan et al., 2009). This study compared more than 100 schools throughout the U.S. that were implementing one of three comprehensive school reform models: Success for All, America's Choice, or Accelerated Schools. There was also a

control group. Students in the SFA portion of the study were followed from kindergarten to second grade. The detailed findings were reported by Correnti (2009), who found an overall effect size of +0.43.

Substantive and Methodological Moderators

Several important demographic and methodological moderators of treatment impacts were identified and explored statistically (see Table 3). Not all coded factors and potential moderators were able to be examined, because of very unequal distributions of studies within moderators, or substantial correlations between moderators and study features.

Research design. Differences in effect sizes between studies that used randomized designs (k = 2, ES = +0.23) and studies that used quasi-experimental designs incorporating matching (k = 15, ES = +0.24) were tested. This difference was not statistically significant.

Evaluator Status. We also compared differences in effect sizes for studies conducted independently from the SFA developers and those conducted in collaboration with SFA. Effect sizes for studies from independent evaluations (ES = +0.21, p < .10) were similar to those from studies conducted with the program developrs (ES = +0.30, p < .10). This difference was not statistically significant.

Duration. Effect sizes were compared for studies at the end of 1 year, 2 years, and 3 or more years. Effect sizes averaged +0.25 after one year, +0.46 after two years, and +0.19 after 3 or more years. Appendix 3 in the online appendix shows year-by-year outcome trends for longitudinal studies, with mean outcomes by year similar to the duration findings.

Race & Ethnicity. Outcomes for samples of mostly African-American students averaged +0.30 (p < .05; *k*=6). In mostly Hispanic samples (*k*=3), effect sizes averaged +0.24 (n.s.). One study included mostly White students, with average effects of +0.63 (p < .05). The remaining ten studies included outcomes of a mix of race and ethnicities, with mean effect sizes of +0.23 (n.s.)

English Learner Status. Impacts were similar for English Learners (ES = +0.13, p < .05), non-English Learners (ES = +0.36, p < .05) and mixed samples (ES = +0.23, p < .05). These differences were not statistically significant.

Achievement Status. Outcomes including all students had a mean effect size of +0.24 (k = 17). Outcomes for low achievers averaged +0.54 (p < .01), significantly higher than outcomes for average/high achievers (ES = +0.07, n.s.), and those for mixed samples (ES=+0.16, n.s.).

Outcome type. Differences in effect sizes across outcome types were also statistically examined. The mean effect size across studies with general reading or comprehension outcomes was +0.20 (n = 90). This contrasted with mean effect sizes across alphabetics outcomes (ES = +0.32, n = 97), and fluency outcomes (ES = +0.14, n = 34). Alphabetics outcomes were significantly higher than fluency outcomes (p < .01).

Discussion

Success for All is a very unusual educational reform program, unique in many ways. It has operated for 33 years with the same basic philosophy and approach, although it has constantly changed its specific components in response to its learnings (Peurach, 2011). Its dissemination has waxed and waned with changing educational policies, SFA served as many as 1500 schools at one time, in 2000-2001. Currently, there are about 500 schools using the full program and another 500 schools using components. In contrast, in two prominent charter networks, KIPP serves 242 schools, and New York's Success Academies serve 45. Also, the program is relatively long-lasting. Data reported by Slavin et al., 2009, indicates that the median SFA school stays in the program for 11 years, and there are several that have used it more than 20 years. At a cost of $117 per student per year (as reported by Quint et al., 2015), SFA is relatively cost-effective (see Borman & Hewes, 2002).

In its long history, Success for All has frequently been evaluated, mostly by third

parties. There were 17 studies that met rigorous inclusion standards. In contrast, the great majority of programs that met the inclusion standards of the What Works Clearinghouse or Evidence for ESSA have been evaluated in just one qualifying study, and very few have been evaluated more than twice.

Across the 17 qualifying U. S. studies, the mean effect size was +0.24 for students in general, and among 8 studies that separately analyzed effects for the lowest achievers, the mean was +0.56. These are important outcomes. As a point of comparison, the mean difference in National Assessment of Educational Progress (NAEP) reading achievement between students who qualify for free lunch and those who do not is approximately an effect size of 0.50 (NCES, 2019). The mean outcomes of Success for All are almost half of this gap, and the outcomes for lowest achievers equal the entire gap.

An important and interesting question for policy and practice is whether SFA works particularly well with sub-populations. The only important factor with sufficient studies to permit subsample analyses was lowest-achieving students (usually students in the lowest 25% of their classes). As noted earlier, the mean effect size for low achievers was +0.54.

It is possible to speculate about what aspects of SFA made the program more effective for lowest achieving students. Low achievers are most likely to receive one-to-one or one-to-small group tutoring, known to have a substantial impact on reading achievement (Neitzel et al., 2020b; Wanzek et al., 2016). Also, there is evidence that cooperative learning, used throughout SFA, is particularly beneficial for low achievers (Slavin, 2017).

The findings of the subgroup analyses with low achievers may be especially important for schools serving large numbers of students who are poor readers. Quint and her colleagues argued that the cost of SFA, which they estimated at $117 per pupil per year, was relatively modest when compared to that of business-as-usual reading programs. In other words, for schools with a high percentage of poor readers, SFA offers a pragmatic alternative supported

by evidence of effectiveness.

The effects of SFA are generally maintained as long as the program remains in operation. In the one study to assess lasting impacts (Borman & Hewes, 2002), outcomes maintain in follow-up as well. This is an unusual finding, and contrasts with the declining impacts over time seen for intensive early tutoring (e.g., Blachman et al., 2014; Pinnell et al., 1994). Beyond SFA itself, this set of findings suggests that a strategy of intensive tutoring and other services followed up with continued interventions to improve classroom instruction to maintain early gains may have more promise than intensive early intervention alone.

The importance of tutoring for struggling readers in the early elementary grades is suggested by the substantially greater short- and long-term impacts of SFA for the lowest-achieving students, who are those most likely to receive tutoring, of course. Another interesting point of comparison also speaks to the importance of tutoring as part of the impact of SFA. Of the four largest evaluations of SFA, three found strong positive impacts. In these, schools were able to provide adequate numbers of tutors to work with most struggling readers in grades 1-3. However, the fourth study, by Quint et al. (2015), took place at the height of the Great Recession (2011-2014). School budgets were severely impacted, and during this study, most schools did not have tutors. This study reported significant positive effects for low achievers, but all outcomes were much smaller than those found in the other studies.

Many phonetic reading programs emphasizing early intervention show substantial positive effects on measures of alphabetics, but not comprehension or general reading. The outcomes of SFA are strongest on measures of alphabetics (ES=+0.32), but are also positive on general reading/comprehension (ES=+0.19), indicating that the program is more than just phonics.

A distressingly common finding in studies of educational programs is that studies carried out by program developers produce much more positive outcomes than do independent

evaluations (Wolf et al., 2020). In the case of Success for All, studies including SFA developers as co-investigators (k=4) do obtain higher effect sizes than do independent studies (k=13) (ES=+0.30 vs. +0.21, respectively), but this difference is not statistically significant. However, this analysis was underpowered, with only 17 studies, so these results must be interpreted with caution.

Policy Importance of Research on Success for All

Attempts to improve the outcomes of education for disadvantaged and at-risk students fall into two types. One focuses on systemwide policies, such as targeted funding, governance, assessment/accountability schemes, standards, and regulations.   These types of strategies are rarely found to be very effective, but they do operate on a very large scale. In contrast, research and development often creates effective approaches, proven to make a meaningful difference in student achievement. However, these proven approaches rarely achieve substantial scale, and if they do, they often do not maintain their effectiveness at scale (see Cohen & Moffitt, 2009, for a discussion of this dilemma).

Success for All is one of very few interventions capable of operating at a scale that is meaningful for policy without losing its effectiveness. At its peak, Success for All operated nationally in more than 1500 schools, and its growth was only curtailed by a shift in federal policies in 2002. Its many evaluations, mostly done by third party evaluators, have found positive outcomes across many locations and over extended periods of time.

In the current policy climate in the U.S., in which evidence of effectiveness is taking on an ever-greater role, Success for All offers one of very few approaches that could, in principle, produce substantial positive outcomes at large scale, and this should have meaning for national policies.

The importance of Success for All for policy and practice is best understood by placing the program in the context of other attempts to substantially improve student achievement in

elementary schools serving many disadvantaged students. A recent review of research on programs for struggling readers in elementary schools by Neitzel et al. (2020b) found that there were just three categories of approaches with substantial and robust evidence of positive outcomes with students scoring in the lowest 25% or 33% of their schools in reading. One was one-to-one or one-to-small group tutoring, by teachers or teaching assistants, with a mean effect size of +0.29. Another was multi-tier whole school/whole school approaches, consisting of Success for All and one other program. The third was whole class Tier 1 programs, mostly using cooperative learning. What these findings imply is that in schools with relatively few students struggling in reading, tutoring may be the best solution for the individuals who are struggling. Even though tutoring is substantially more expensive per student than Success for All, in a school with few struggling readers, it may not be sensible to intervene with all students.

On the other hand, when most students need intervention in reading, it is not sensible or cost-effective to solve the problem with tutoring alone. In the U.S., the average large urban school district has only 28% of fourth graders scoring "proficient" or better on the National Assessment of Educational Progress (NCES, 2019), and in cities such as Dallas, Milwaukee, Baltimore, Cleveland, and Detroit, fewer than 15% of students in the entire district score at or above "proficient." In such districts, and in individual low-performing schools even in higher-performing districts, trying to reach high levels of proficiency through tutoring alone would be prohibitively expensive.

The findings of the evaluations of Success for All have particular importance for special education policies. The structure of SFA adheres closely to the concept of Response to Intervention (RTI). SFA emphasizes professional development, coaching, and extensive programming to improve outcomes of Tier 1 classroom instruction, which is then followed up by closely coordinated Tier 2 (small-group tutoring) or Tier 3 (one-to-one tutoring) for

students who need it. Longitudinal research found substantial and lasting impacts on the achievement of the lowest achievers, and on reductions in assignment to special education as well as retentions in grade (Borman & Hewes, 2002).

Beyond the program itself, the research on Success for All, as applied to low-achieving students, illustrates that the educational problems of low-achieving students are fundamentally solvable. Perhaps someday there will be many approaches like Success for All, each of which is capable of improving student achievement on a substantial scale. Research on Success for All suggests that disadvantaged students and struggling readers could be learning to read at significantly higher levels than they do today, and that substantial improvement can be brought about at scale. The knowledge that large-scale improvement is possible should lead to policies that both disseminate existing proven approaches, and invest in research and development to further increase the effectiveness and replicability of programs that can reliably produce important improvements in reading for disadvantaged and low-achieving readers.

References

*A double asterisk indicates studies included in the main meta-analysis (final reports).

*A single asterisk indicate studies included in the exploratory meta-analysis (interim reports).

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive Reading Remediation in grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology*, *106*(1), 46–57.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009*). Introduction to meta-analysis*. Chichester, England: John Wiley & Sons, Ltd.

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24*(4), 243–266. https://doi.org/10.3102/01623737024004243

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230. https://doi.org/10.3102/00346543073002125

**Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44*(3), 701–731. https://doi.org/10.3102/0002831207306743

*Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005a). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal*, *42*(4), 673–696.

*Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005b). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis*, *27*(1), 1–22. https://doi.org/10.3102/01623737027001001

**Chambers, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2005). *Enhancing Success for All for Hispanic students: Effects on beginning reading achievement*. Baltimore, MD: Success for All Foundation. http://eric.ed.gov/?id=ED485350

Chambers, B., Cheung, A., & Slavin, R. (2016). Literacy and language outcomes of balanced and developmental-constructivist approaches to early childhood education: A systematic review. *Educational Research Review 18,* 88-111.

Cohen, D. K., & Moffitt, S. L. (2009). *The ordeal of equality: Did federal regulation fix the schools?* Cambridge, MA: Harvard University Press.

**Correnti, R. (2009). *Examining CSR program effects on student achievement: Causal explanation through examination of implementation rates and student mobility*. Paper presented at the 2nd annual conference of the Society for Research on Educational Effectiveness, Washington, DC, March, 2009.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934-945. doi: 10.1037/0012-1649.33.6.934

**Datnow, A., Stringfield, S., Borman, G., Rachuba, L., & Castellano, M. (2001). *Comprehensive School Reform in Culturally and Linguistically Diverse Contexts: Implementation and Outcomes from a 4-Year Study*. Santa Cruz, CA: Center for Research on Education, Diversity, and Excellence.

Fuchs, D., & Fuchs, L. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41* (1), 92-128. https://doi.org/10.1598/RRQ.41.1.4

Good, T., & Brophy, J. (2018). *Looking in classrooms (10ᵗʰ ed.).* Boston: Allyn & Bacon.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Lesnick, J., Goerge, R., Smithgall, C., & Gwynne J. (2010). *Reading on grade level in third grade: How is it related to high school performance and college enrollment?* Chicago: Chapin Hall at the University of Chicago.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

**Livingston, M., & Flaherty, J. (1997). *Effects of Success for All on reading achievement in California schools.* Los Alamitos, CA: WestEd.

Madden, N. A., & Slavin, R. E. (2017). Evaluations of technology-assisted small-group tutoring for struggling readers. *Reading & Writing Quarterly,* 1-8. http://dx.doi.org/10.1080/10573569.2016.1255577

**Madden, N.A., Slavin, R.E., Karweit, N.L., Dolan, L.J., Wasik, B.A. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal, 30*, 123-148. https://doi.org/10.3102/00028312030001123

**Muñoz, M.A., Dossett, D., & Judy-Gullans, K. (2004). Educating students placed at risk: Evaluating the impct of Success for All in urban settings. *Journal of Education for Students Placed at Risk, 9*(3), 261-277. https://doi.org/10.1207/s15327671espr0903_3

National Center for Education Statistics (NCES) (2019). *The condition of education 2019.* Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144

National Reading Panel (NRP) (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Pub. No. 00-4754). Retrieved from http://www.nichd.nih.gov/publications/pubs/nrp/pages/report.aspx

Neitzel, A., Cheung, A., Xie, C., Zhuang, T.T., & Slavin, R. (2020a). Data archive for "Success for All: A Quantitative Synthesis of U. S. Evaluations." Towson, MD: Center for Research and Reform in Education (CRRE), Johns Hopkins University. Retrieved from https://github.com/aj-neitzel/Success-for-All-A-Quantitative-Synthesis-of-U.-S.-Evaluations

Neitzel, A., Lake, C., Pellegrini, M., & Slavin, R. (2020b). *Effective programs for struggling readers: A best-evidence synthesis.* Manuscript submitted for publication. Also available at www.bestevidence.org.

**Nunnery, J.A., Slavin, R., Ross, S., Smith, L., Hunter, P., & Stubbs, J. (1996*). An assessment of Success for All program component configuration effects on the reading achievement of at-risk first grade students.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

OECD (2019). *PISA 2018 Technical Report*. Paris, France: OECD Publishing.

Peurach, D.J. (2011). *Seeing complexity in public education: Problems, possibilities, and Success for All*. New York, New York: Oxford University Press.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research, 90* (1), 24-46.

Pinnell, G. S., Lyons, C. A., DeFord, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high risk first graders. *Reading Research Quarterly, 29,* 8–38.

Pustejovsky, J. (2020). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections* (Version R package version 0.4.1) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

**Quint, J., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015*). Scaling up the Success for All model of school reform: Final report from the Investing in Innovation (i3) evaluation.* New York: MDRC.

*Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J., & Zhu, P. (2013). *The Success for All model of school reform: Early findings from the Investing in Innovation (i3) scale-up*. MDRC.

*Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J., & Zhu, P. (2014). *The Success for All model of school reform: Interim findings from the Investing in Innovation (i3) scale-up*. MDRC.

R Core Team. (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

**Ross, S.M., & Casey, J. (1998). *Success for All evaluation, 1997-98 Tigard-Tualatin School District.* Memphis: University of Memphis, Center for Research on Educational Policy.

**Ross, S.M., Nunnery, J.A., & Smith, L.J. (1996a). *Evaluation of Title I reading programs: Amphitheater Public Schools Year 1: 1995-1996*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

**Ross, S.M., Nunnery, J., Smith, L.J., & Lewis, T. (1997b). *An evaluation of Roots & Wings in Memphis City Schools*. Memphis: University of Memphis, Center for Research in Educational Policy.

**Ross, S.M., Smith, L.J., & Bond, C. (1994a). *An evaluation of the Success for All program in Montgomery, Alabama schools*. Memphis, TN: University of Memphis, Center for Research on Educational Policy.

**Ross, S.M., Smith, L.J., & Casey, J.P. (1995). *Final Report: 1994-1995 Success for All program in Fort Wayne, Indiana.* Memphis, TN: University of Memphis, Center for Research in Educational Policy

**Ross, S.M., Smith, L.J., & Casey, J.P. (1997c). Preventing early school failure: Impacts of Success for All on standardized tests outcomes, minority group performance, and school effectiveness. *Journal of Education for Students Placed at Risk, 2*(1), 29-53. https://doi.org/10.1207/s15327671espr0201_4

*Ross, S.M., Smith, L.J., Casey, J.P., Johnson, B., & Bond, C. (1994b). *Using Success for All to restructure elementary schools: A tale of four cities*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 373456)

**Ross, S.M., Smith, L.J., Lewis, T., & Nunnery, J. (1996). *1995-96 evaluation of Roots & Wings in Memphis City Schools.* Memphis: University of Memphis, Center for Research in Educational Policy.

Rowan, B., Correnti, R., Miller, R., & Camburn, E. (2009). *School improvement by design: Lessons from a study of comprehensive school reform programs*. At http://www.cpre.org/school-improvement-design-lessons-study-comprehensive-school-reform-programs.

Shaywitz, S. E., & Shaywitz, J. (2020). *Overcoming dyslexia (2nd ed.).* New York: Penguin Random House.

Slavin, R.E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57*, 347-350. https://doi.org/10.3102/00346543057003293

Slavin, R. E. (2017). Instruction based on cooperative learning. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction*. New York: Routledge.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391–1466. https://doi.org/10.3102/0034654309341374

Slavin, R.E., Madden, N.A., Chambers, B., & Haxby, B. (2009). *Two million children: Success for All*. Thousand Oaks, CA: Corwin.

**Slavin, R.E., Madden, N.A., Dolan, L.J., & Wasik, B.A. (1993*). Success for All in the Baltimore City Public Schools: Year 6 Re*port*. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

**Slavin, R.E., Madden, N.A., Dolan, L.J., Wasik, B.A., Ross, S.M., & Smith, L.J. (1994). *Success for All: Longitudinal effects of systemic school-by-school reform in seven districts*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

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

Stevens, R. J., Madden, N. A., Slavin, R. E., & Farnish, A. M. (1987). Cooperative Integrated Reading and Composition: Two field experiments. *Reading Research Quarterly, 22*, 433-454. doi: 10.2307/747701

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the **metafor** package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

**Wang, W., & Ross, S.M. (1999). *Results for Success for All program, Alhambra (AZ) School District*. Memphis, TN: University of Memphis, Center for Research on Educational Policy.

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review 28(*3), 551-576. https://doi.org/10.1007/s10648-015-9321-7

Wolf, R., Morrison, J.M., Inns, A., Slavin, R. E., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness.* DOI: 10.1080/19345747.2020.1726537

What Works Clearinghouse. (2020). *Standards Handbook (Version 4.1)*. Institute of Education Sciences, US Department of Education.

What Works Clearinghouse. (2014). *Review protocol for beginning reading interventions version 3.0*. Institute of Education Sciences, US Department of Education.

Table 1: Meta-regression results.

	Reference	Coefficient	SE	t	df	p
Null Model						
Intercept		0.10	0.06	1.59	9.42	0.146
Meta-Regression						
Success for All vs. Control (Intercept)		0.24	0.08	3.07	7.24	0.017
Randomized Studies	Quasi-Experiments	-0.05	0.18	-0.27	3.14	0.804
Independent Evaluations	Not Independent Evaluations	-0.07	0.13	-0.48	4.39	0.653
1 year studies	3+ year studies	0.06	0.14	0.44	7.92	0.670
2 year studies		0.27	0.11	2.43	1.91	0.141
Black Students	Mix of Students	0.08	0.20	0.42	5.54	0.687
White Students		0.41	0.23	1.74	3.79	0.162
Hispanic Students		0.06	0.27	0.21	2.75	0.846
No EL Students	Mix of Language Status Students	0.10	0.08	1.13	2.39	0.358
EL Students		0.04	0.07	0.59	2.06	0.615
General Reading/ Comprehension Outcomes	Fluency Outcomes	0.05	0.06	0.83	5.05	0.443
Alphabetics Outcomes		0.18	0.03	6.93	4.97	0.001
Low Achievers	Moderate/High Achievers	0.46	0.15	3.19	4.33	0.030
Mix of Students		0.09	0.05	2.03	2.68	0.146

Note. SE=standard error; df=degrees of freedom

Table 2: Features and Summary of Outcomes of Included Studies

Study	Design	Evaluator	Sample	Sample Description	n	Outcome	Duration	Grade	Study ES	Low Achiever ES
Quint et al. (2015)	CR	Ind.	37 schools 1635 students	Five school districts, mostly in or on outskirts of large or midsize cities in the Northeast, South, and West 12% W, 18%AA, 88%FRL, 24%ELL, 66%H	28	GR/C, Al	3 years	K-2	+0.08	+0.18
Borman et al. (2007)	CR		35 schools 2108 students	Title I schools throughout the U.S. 72%FRL, 56%AA, 30%W, 10%H	3	GR/C, Al	3 years	K-2	+0.25	
Correnti (2009)	CQE	Ind.	115 schools 3783 students	High-poverty schools in 17 states 69%FRL, 52%AA, 21%W, 18%H	1	GR/C	3 years	K-2	+0.43	
Nunnery et al. (1996)	CQE		67 schools 2060 students	High-poverty schools in Houston, TX 78%FRL, 54%H, 38%AA	3	GR/C	1 year	1st	+0.19	
Ross et al. (1996b)	CQE	Ind.	12 schools 781 students	Memphis, TN	4	GR/C, Al, Fl	1 year	1st	+0.01	
Slavin et al. (1993)	CQE		10 schools 1495 students	African-American students in high-poverty schools in Baltimore, MD	59	GR/C, Al, Fl	3 years	preK-1	+0.59	+1.17
							4 years	preK-2	+0.29	+0.91

Study	Design	Evaluator	Sample	Sample Description	n	Outcome	Duration	Grade	Study ES	Low Achiever ES
							5 years	preK-3	+0.41	+1.29
							6 years	preK-4	+0.41	+0.78
								K-5	+0.46	+1.01
Chambers et al. (2005)	CQE		8 schools 577 students	Mostly Hispanic communities in the U.S.	8	GR/C, Al	1 year	K	+0.28	
								1st	+0.32	
Ross & Casey (1998a)	CQE	Ind.	8 schools 356 students	High-poverty schools in Ft. Wayne, IN 75%FRL, 45%minority	8	GR/C, Al, Fl	2 years	K-1	+0.26	+0.34
Datnow et al. (2001)	CQE	Ind.	6 schools 398 students	Diverse students in Miami, FL	2	GR/C	4 years	1-4	+0.11	
Livingston & Flaherty (1997)	CQE	Ind.	6 schools 828 students	High-poverty multilingual schools in Modesto and Riverside, CA.	12	GR/C	2 years	K-1	+0.65	
							3 years	K-2	+0.40	
							4 years	K-3	+0.12	
Munoz & Dossett (2004)	CQE	Ind.	6 schools 349 students	High-poverty schools in Louisville, KY	1	GR/C	3 years	1-3	+0.15	
Ross et al. (1996a)	CQE	Ind.	5 schools 428 students	Tucson, Arizona	16	GR/C, Al, Fl	1 year	1st	+0.41	+0.51

Study					n			ES	Grade	ES	ES
Ross & Casey (1998b)	CQE	Ind.	4 schools 581 students	Suburban schools in Portland, OR. 15%minority	16	GR/C, Fl	Al,	1 year	K	+0.16	+0.36
									1st	-0.02	-0.18
Ross et al. (1994a)	CQE	Ind.	4 schools 179 students	African-American students in high-poverty schools in Montgomery, AL	8	GR/C, Fl	Al,	2 years	1-2	+0.58	+1.01
Ross et al. (1995)	CQE	Ind.	4 schools 257 students	Title I schools in Ft. Wayne, IN	20	GR/C, Fl	Al,	3 years	K-2	+0.10	+0.56
								4 years	K-3	-0.10	
									1-4	0.00	+0.29
Ross et al. (1997)	CQE	Ind.	4 schools 291 students	A medium-size midwestern city	24	GR/C, Fl	Al,	2 years	K-1	+0.28	+0.86
								3 years	K-2	+0.16	
									1-3	+0.02	
Wang & Ross (1999)	CQE	Ind.	4 schools 340 students	High poverty African-American schools in Little Rock, AK	8	GR/C, Fl	Al,	1 year	1st	+0.24	
									2nd	-0.05	

Note. CR = Cluster randomized; CQE = Cluster quasi-experiment; Ind. = Independent; W = White; AA = African American; FRL = Free/Reduced lunch; ELL = English Language Learner; H = Hispanic; GR/C = General Reading/Comprehension; AL = Alphabetics; FL = Fluency; n = number of effect sizes; ES = effect size

Table 3: Substantive and methodological moderators

Moderator	Level	k	n	ES	SE	t	df	p
Research Design	Randomized	2	31	+0.20	0.14	1.44	1.51	0.322
	Quasi-Experiment	15	190	+0.25	0.09	2.71	6.36	0.033
Duration	1 year	6	55	+0.25	0.09	2.91	4.88	0.035
	2 years	4	32	+0.46	0.15	3.11	5.19	0.025
	3+ years	9	134	+0.19	0.10	1.77	5.12	0.135
Race	Black	6	85	+0.28	0.10	2.80	4.27	0.046
	Hispanic	3	34	+0.25	0.24	1.06	3.68	0.355
	White	1	4	+0.60	0.21	2.92	4.14	0.042
	Mix	10	98	+0.19	0.14	1.37	7.79	0.210
Language Learner Status	EL	4	18	+0.27	0.10	2.80	4.55	0.042
	Not EL	4	12	+0.33	0.10	3.41	4.70	0.021
	Mix	15	191	+0.23	0.08	2.88	7.74	0.021
Outcome Type	General Reading/ Comprehension	17	90	+0.19	0.08	2.52	7.51	0.038
	Alphabetics	12	97	+0.32	0.09	3.50	7.44	0.009
	Fluency	9	34	+0.14	0.08	1.71	7.03	0.132
Achievement Status	Low Achiever	8	60	+0.54	0.15	3.69	6.16	0.010
	Average/High Achiever	8	60	+0.07	0.07	1.05	5.36	0.338
	Mix	14	101	+0.16	0.08	2.00	8.23	0.080
Evaluator Status	Independent Evaluator	13	148	+0.22	0.10	2.18	9.50	0.056
	Not Independent Evaluator	4	73	+0.28	0.10	2.80	3.02	0.067

Note. k = number of studies; n = number of outcomes; ES = effect size; SE = standard error; df = degrees of freedom. Mean effect sizes for each moderator category were calculated by estimated a model including the same covariates as those shown in Table 1 without an intercept, with the moderator included as a categorical variable.

Figure 1: Theory of Action for Success for All
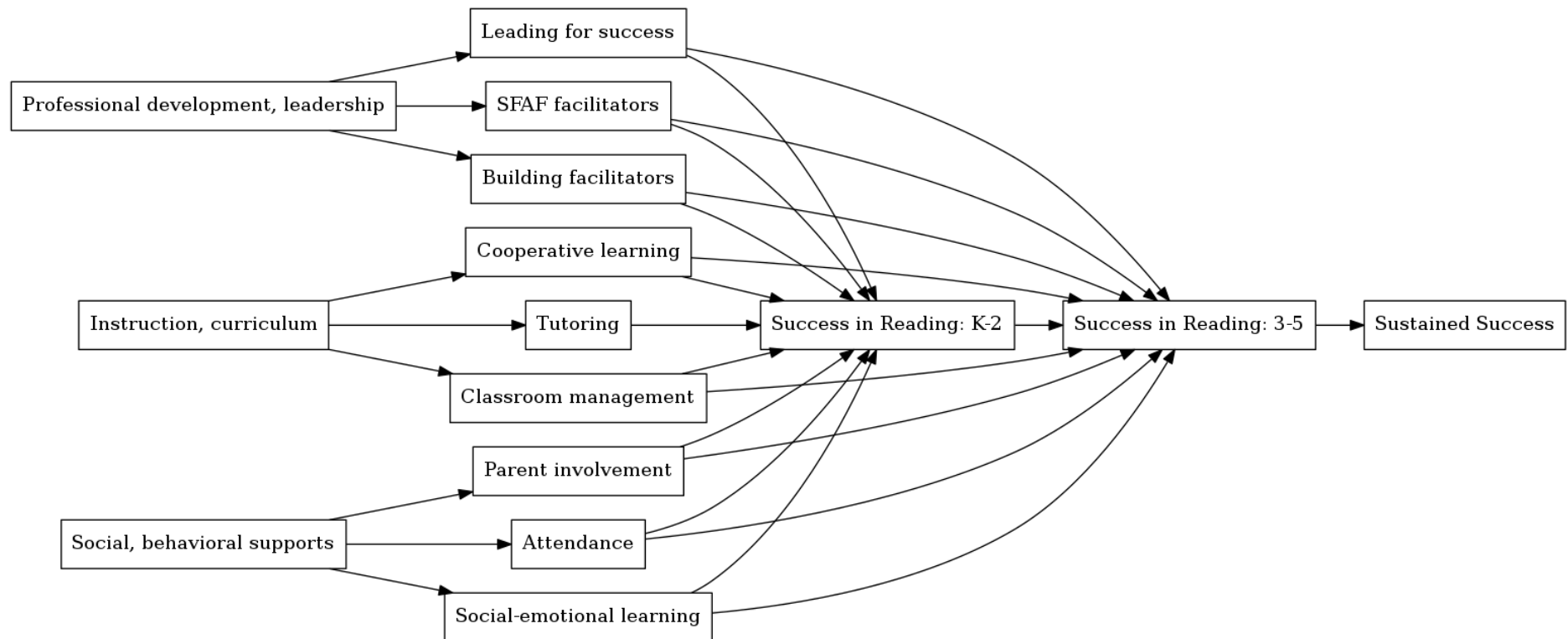
Figure 2: Flow Chart of Study Selection
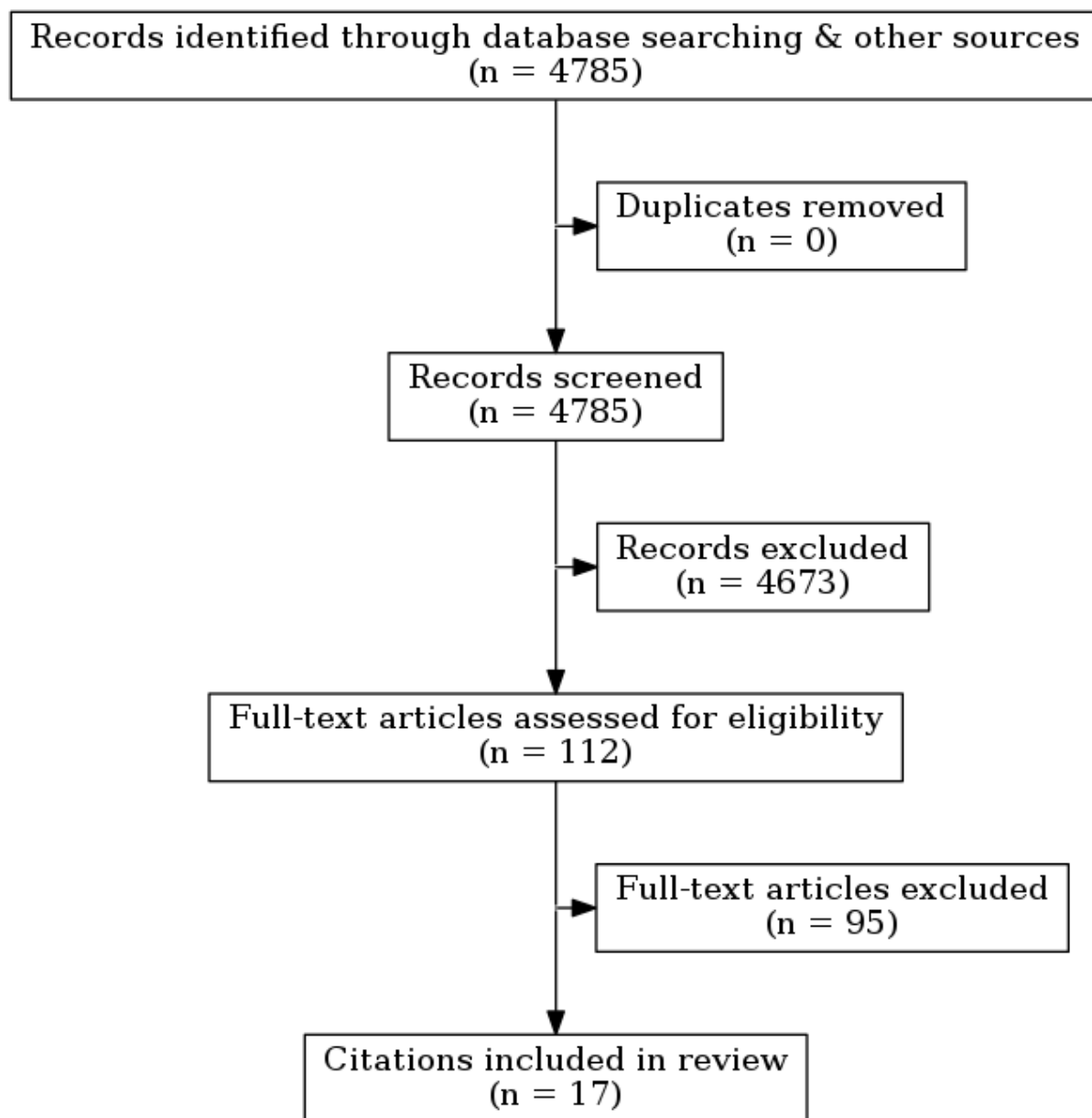
Appendix 1: Selected Studies Excluded by the Review

Reference	Reasons for excluding
Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). The national randomized field trial of Success for All: Second-year outcomes. American Educational Research Journal, 42(4), 673–696.	Interim report
Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). Success for All: First-year results from the national randomized field trial. Educational Evaluation and Policy Analysis, 27(1), 1–22. https://doi.org/10.3102/01623737027001001	Interim report
Calderon, M. (2001). Success for All in Mexico. Success for All: Research and reform in elementary education. Mahwah, NJ: Erlbaum.	No appropriate data.
Calderon, M., August, D., Slavin, R. E., Duran, D., Madden, N. A., & Cheung, A. (2004). The evaluation of a bilingual transition program for Success for All. Baltimore, MD: Center for Research on the Education of Students Placed at Risk.	Upper elementary
Casey, J., Smith, L.J., Ross, S.M. (1994). Formative evaluation of new Success for All schools, Memphis, Tennessee, 1993-1994. Memphis: University of Memphis, Center for Research in Educational Policy.	Insufficient schools
Center, Freeman, & Robertson. (1998). An evaluation of Schoolwide Early Language and Literacy (SWELL) in six disadvantaged schools. International Journal of Disability, Development and Education, 45, 143-172.	No pretest.
Center, Y. & Freeman, L. (1997). A trial evaluation of SWELL (Schoolwide Early Language and Literacy Program) for at-risk and disadvantaged children. International journal of Disability, Development and Education, 44, 21-39. https://doi.org/10.1080/0156655970440103	Non-U.S.
Center, Y., Freeman, L., & Robertson, G. (2001a). The relative effect of a code-oriented and meaning-oriented early literacy program on regular and low-progress Australian students in Year 1 classroom using Reading Recovery. International journal of Disability, Development and Education, 48(2), 207-232. https://doi.org/10.1080/10349120120053676	Non-U.S.

Center, Y., Freeman, L., & Robertson, G. (2001b). A longitudinal evaluation of the Schoolwide Early Language and Literacy Program (SWELL). In R.E. Slavin & N.A. Madden (Eds.), Success for All: Research and reform in elementary education. Mahwah, NJ: Erlbaum.	Non-U.S.
Chambers, B., Abrami, P.C., & Morrison, S. (2001). Can Success for All succeed in Canada? In R.E. Slavin (Ed.), Success for All: Research and reform in elementary education. Mahwah, NJ: Erlbaum.	Non-U.S.
Chambers, B., Abrami, P.C., Tucker, B.J., Slavin, R.E., Madden, N.A., Cheung, A., & Gifford, R. (2008). Computer assisted tutoring in Success for All: Reading outcomes for first grade. Journal of Research on Effective Education, 1 (2), 120-137.	Inadequate comparison group
Chambers, B., Cheung, A., Madden, N. A., Slavin, R. E., & Gifford, R. (2006). Achievement effects of embedded multimedia in a Success for All reading program. Journal of Educational Psychology, 98(1), 232-237.	Inadequate comparison group
Chambers, B., Slavin, R. E., Madden, N. A., Abrami, P. C., Karanzalis, M., & Gifford, R. (2010). Small-group computer-assisted tutoring to improve reading outcomes for struggling first and second graders. Elementary School Journal, 111 (4), 625-640.	Inadequate comparison group
Chambers, B., Slavin, R.E., Madden, N.A., Abrami, P.C., Tucker, B.J., Cheung, A., & Gifford, R. (2008a). Technology Infusion in Success for All: Reading Outcomes for First Graders. The Elementary School Journal, 109(1), 1-15.	Inadequate comparison group
Dianda, M. R., & Flaherty, J. F. (1995). Effects of Success for All on the reading achievement of first graders in California bilingual programs.	Interim report
Education Quality Institute (2002). Progress and options regarding the implementation of Direct Instruction and Success for All in Toledo Public Schools. Toledo, OH: Education Quality Institute.	Inadequate comparison group
Hanselman, P., & Borman, G. D. (2013). The impacts of Success For All on reading achievement in grades 3–5: Does intervening during the later elementary grades produce the same benefits as intervening early? Educational Evaluation and Policy Analysis, 35(2), 237–251.	Upper elementary
Harris, A., Hopkins D., & Wordsworth, J. (2001). The implementation and impact of Success for All in English	No appropriate data.

schools. Success for All: Research and reform in elementary education. Mahwah, NJ: Erlbaum.	
Hertz-Lazarowitz, R. (2001). Success for All: A model for advancing Arabs and Jews in Israel. Success for All: Research and reform in elementary education. Mahwah, NJ: Erlbaum.	No pretest.
Hopkins, D., Youngman, M., Harris, A., & Wordsworth, J. (1999). Evaluation of the initial effects and implementation of Success for All in England. Journal of Research in Reading, 22 (3), 257-270.	No appropriate data.
Hurley, E., Chamberlain, A., Slavin, R.E., & Madden, N.A. (2001). Effects of Success for All on TAAS Reading Scores: A Texas statewide evaluation. Phi Delta Kappan, 82 (10), 750-756.	Inadequate comparison group
Jones, E.M., Gottfredson, G.D., Gottfredson, D.C. (1997). Success for Some: An Evaluation of a Success for All Program. Evaluation Review, 21(6), 643-670.	Insufficient schools
Madden, N.A., Slavin, R.E., Karweit, N.L., Dolan, L.J., Wasik, B.A. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. American Educational Research Journal, 30, 123-148. https://doi.org/10.3102/00028312030001123	Interim report
Miller, S., Biggart, A., Sloan, S., & O'Hare, L. (2017). Success for All: Evaluation report and executive summary July 2017. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Success_for_All_Evaluation_Report.pdf	Non-U.S.
New American Schools. (1997). Working towards Excellence: Results from Schools Implementing New American Schools Designs. Arlington, VA: New American Schools.	No appropriate data.
Nunnery, J., Slavin, R.E., Madden, N.A., Ross, S., Smith, L., Hunter, P., Stubbs, J. (1997). Effects of full and partial implementation of Success For All on student reading achievement in English and Spanish. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.	Measure in Spanish
Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J., & Zhu, P. (2013). The Success for All model of school reform: Early findings from the Investing in Innovation (i3) scale-up. MDRC.	Interim report
Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J.,	Interim report

& Zhu, P. (2014). The Success for All model of school reform: Interim findings from the Investing in Innovation (i3) scale-up. MDRC.	
Ross, S. M., Smith, L. J., Johnson, J. P., & Bond, C. (1994b). Using Success For All to restructure elementary schools: A tale of four cities. Annual Meeting. American Educational Research Associatio, New Orleans, LA.	Interim report
Ross, S. M., Wang, L. W., Sanders, W. L., Wright, S. P., & Stringfield, S. (1999). Two- and three-year achievement results on the Tennessee Value-Added Assessment System for restructuring schools in Memphis. Center for Research in Educational Policy.	No pretest.
Ross, S.M., Nunnery, J., Smith, L.J., & Lewis, T. (1997b). An evaluation of Roots & Wings in Memphis City Schools. Memphis: University of Memphis, Center for Research in Educational Policy.	Manuscript unavailable.
Ross, S.M., Sanders, W.L., Wright, S.P. (2000). Fourth Year Achievement Results on the Tennessee Value-Added Assessment System for Restructuring Schools in Memphis. Memphis: University of Memphis, Center for Research in Educational Policy.	No measure for reading achievement.
Ross, S.M., Smith, L.J., & Casey, J. (1992). Final report: 1991-92 Success for All program in Caldwell, Idaho. Memphis, TN: Memphis State University.	Insufficient schools
Ross, S.M., Smith, L.J., & Casey, J. (1994). Final report: 1993-94 Caldwell, Idaho. SFA results. Memphis, TN: The University of Memphis. Center for Research in Educational Policy.	Insufficient schools
Ross, S.M., Smith, L.J., Casey, J.P. (1997a). Final report: 1996-97 Success for All program in Clark County, Georgia. Memphis, TN: University of Memphis, Center for Research on Educational Policy.	Insufficient schools
Ross, S.M., Smith, L.J., Casey, J.P., Johnson, B., Bond, C. (1994). Using Success For All to restructure elementary schools: A tale of four cities. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.	Insufficient schools
Ross, S.M., Smith, L.J., Crawford, A., Nunnery, J., Eck, L., Lohr, L., & Faulks, A. (1991). Final report: 1991 Success for All program in Memphis. Memphis, TN: University of Memphis, Center for Research in Educational Policy.	Insufficient schools
Ross, S.M., Smith, L.J., Nunnery, J.A. (1998). The relationship of	No pretest.

program implementation quality and student achievement. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.	
Rowan, B., & Miller, R. (2007). Organizational strategies for promoting instructional change. American Educational Research Journal, 44, 252-297.	No appropriate data.
Rowan, B., Correnti, R., Miller, R. J., & Camburn, E. M. (2009). School improvement by design: Lessons from a study of comprehensive school reform programs. Ann Arbor, MI: Consortium for Policy Research in Education.	No appropriate data.
Sanders, W., Wright, S., Ross, S., & Wang, W. (2000). Value-added achievement results for three cohorts of Roots and Wings schools in Memphis: 1995-1999 outcomes. Memphis: University of Memphis, Center for Research in Educational Policy.	No measure for reading achievement.
Slavin, R. E., Madden, N. A., Karweit, N. L., Livermon, B. J., & Dolan, L. (1989). Can every child learn? An evaluation of "Success for All" in an urban school. Journal of Negro Education, 58, 357-366. doi: 10.2307/2295668	Interim report
Slavin, R.E. & Madden, N.A. (1994). Implementing Success for All in Philadelphia Public Schools (final report to the Pew Charitable Trusts). Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Manuscript unavailable.
Slavin, R.E. (1993). Success for All in the Philadelphia public schools: 1991-92 evaluation report. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Manuscript unavailable.
Slavin, R.E., Madden, N.A. (1991). Success for All at Buckingham Elementary: Second year evaluation. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Insufficient schools
Slavin, R.E., Madden, N.A. (1998). Success for All/Exito Para Todos: Effects on the reading achievement of students acquiring English. Report No. 19. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.	Insufficient schools
Slavin, R.E., Madden, N.A., Dolan, L.J., Wasik, B.A., Ross, S.M., & Smith, L.J. (1994). Success for All: Longitudinal effects of systemic school-by-school reform in seven districts. Paper	No pretest.

presented at the annual conference of the American Educational Research Association, New Orleans, LA.	
Slavin, R.E., Madden, N.A., Karweit, N. L., Dolan, L.J., & Wasik, B.A. (1990). Success for All: Second year report. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Interim report
Slavin, R.E., Madden, N.A., Karweit, N.L., Dolan, L.J., Wasik, B.A. (1993). Success For All in the Baltimore City public schools: Year 5 Report. Baltimore: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Interim report
Slavin, R.E., Madden, N.A., Karweit, N.L., Livermon, B.J., Dolan, L.J. (1990). Success For All: First-year outcomes of a comprehensive plan for reforming urban education. American Educational Research Journal, 27(2), 255-278. https://doi.org/10.3102/00028312027002255	Interim report
Slavin, R.E., Yampolsky, R. (1991). Success For All: Effects on language minority students. Report No. 14. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Disadvantaged Students.	Insufficient schools
Slavin, R.E., Yampolsky, R. (1992). Effects of Success for All on students with limited English proficiency: A three-year evaluation. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Insufficient schools
Smith, L.J., & Ross, S.M. (1992). 1991-1992 Fort Wayne, Indiana, SFA Results. Memphis, TN: Memphis State University, Center for Research in Educational Policy.	No pretest.
Smith, L.J., Ross, S.M., & Casey, J.P. (1994b). Special education analyses for Success for All in four cities. Memphis: University of Memphis, Center for Research in Educational Policy.	Insufficient schools
Smith, L.J., Ross, S.M., Casey, J.P. (1994). Final Report: 1993-94 Success for All Program in Memphis, Tennessee. Memphis, TN: University of Memphis, Center for Research in Educational Policy.	Insufficient schools
Smith, L.J., Ross, S.M., Johnson, B., Casey, J. (1993). 1992-1993 Memphis, Tennessee SFA results. Final report. Memphis, TN: University of Memphis, Center for Research in Educational Policy.	Insufficient schools
Smith, L.J., Ross, S.M., Johnson, B., Casey, J. (1993). 1992-1993 Memphis, Tennessee SFA results. Final report. Memphis, TN: University of Memphis, Center for Research in Educational	Insufficient schools

Policy.	
Tracey, L., Chambers, B., Slavin, R., & Madden, N. (2011). Success for All in the UK: Results from the second year of a national evaluation. Heslington, UK: Institute for Effective Education, University of York.	Non-U.S.
Tracey, L., Chambers, B., Slavin, R., Hanley, P., & Cheung, A. (2014). Success for All in England: Results from the third year of a national evaluation. SAGE Open, 4(3), 1-10. doi: 10.1177/2158244014547031.	Non-U.S.
Urdegar, S.M. (1998). Evaluation of the Success For All Program 1997-98. Miami, FL: Miami-Dade Public Schools, Office of Evaluation Research.	No pretest.
Wang, W., Ross, S.M. (1999b). Results for Success for All program, Alhambra (AZ) School District. Memphis, TN: University of Memphis, Center for Research on Educational Policy.	Insufficient schools
Wasik, B.A., Slavin, R.E. (1993). Success for All at Pepperhill Elementary School: 1992 evaluation. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.	Insufficient schools

Appendix 2: Details of studies and findings

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
Quint et al. (2015)	CR 3 years	37 schools 1635 students	Five school districts, mostly in or on outskirts of large or midsize cities in the Northeast, South, and West 12% W, 18%AA, 88%FRL, 24%ELL, 66%H	K-2	African-American	TOWRE	+0.16	+0.08
						WJ LWID	+0.09	
						WJ PC	+0.12	
						WJ WA	+0.10	
					Hispanic	TOWRE	-0.03	
						WJ LWID	+0.02	
						WJ PC	-0.01	
						WJ WA	+0.06	
					White	TOWRE	+0.02	
						WJ LWID	+0.46	
						WJ PC	+0.57	
						WJ WA	+0.32	
					ELL	TOWRE	+0.09	
						WJ LWID	+0.01	
						WJ PC	-0.06	
						WJ WA	+0.10	
					Non-ELL	TOWRE	+0.08	
						WJ LWID	+0.09	
						WJ PC	+0.04	
						WJ WA	+0.20	
					Average/high achiever	TOWRE	-0.03	
						WJ LWID	-0.02	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						WJ PC	-0.06	
						WJ WA	+0.07	
						TOWRE	+0.19	
					Low achiever	WJ LWID	+0.17	
						WJ PC	+0.14	
						WJ WA	+0.23	
Borman et al. (2007)	CR 3 years	35 schools 2108 students	Title I schools throughout the U.S. 72%FRL, 56%AA, 30%W, 10%H	K-2		WRMT LWID	+0.22	+0.25
						WRMT PC	+0.21	
						WRMT WA	+0.33	
Correnti (2009)	CQE 3 years	115 schools 3783 students	High-poverty schools in 17 states 69%FRL, 52%AA, 21%W, 18%H	K-2		Terra Nova	+0.43	+0.43
Nunnery et al. (1996)	CQE 1 year	67 schools 2060 students	High-poverty schools in Houston, TX 78%FRL, 54%H, 38%AA	1st	Full model	Woodcock + DORT	+0.31	+0.19
					Reading-only		+0.19	
					Reading + tutoring		+0.13	
Ross et al. (1996b)	CQE 1 year	12 schools 781 students	Memphis, TN	1st		Oral Reading	-0.09	+0.01
						PC	-0.05	
						Word Attack	+0.31	
						Word ID	-0.12	
Slavin et al. (1993)	CQE 3-6 years	10 schools 1495 students	African-American students in high-poverty schools in Baltimore, MD	preK-1	All students	CTBS RC	+0.99	+0.44
						CTBS RV	+0.64	
						CTBS TL	+0.52	
						DORT	+0.40	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						WJ PC	+0.50	
						WJ WA	+0.67	
						WJ WID	+0.48	
						DORT	+0.69	
					Low achiever	WJ PC	+0.92	
						WJ WA	+2.66	
						WJ WID	+0.40	
						CTBS RC	+0.05	
						CTBS RV	+0.17	
						CTBS TL	+0.33	
					All students	DORT	+0.24	
						WJ PC	+0.31	
				preK-2		WJ WA	+0.51	
						WJ WID	+0.40	
						DORT	+0.80	
					Low achiever	WJ PC	+0.42	
						WJ WA	+1.58	
						WJ WID	+0.83	
						CTBS RC	+0.25	
						CTBS RV	+0.10	
				preK-3	All students	CTBS TL	+0.18	
						DORT	+0.53	
						WJ PC	+0.55	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						WJ WA	+0.66	
						WJ WID	+0.55	
					Low achiever	DORT	+1.42	
						WJ PC	+1.09	
						WJ WA	+1.50	
						WJ WID	+1.16	
				preK-4	All students	CTBS RC	+0.34	
						CTBS RV	+0.04	
						CTBS TL	+0.39	
						Gray RC	+0.33	
						Gray RP	+0.51	
						WJ PC	+0.47	
						WJ WA	+0.56	
						WJ WID	+0.59	
					Low achiever	Gray RC	+0.51	
						Gray RP	+0.62	
						WJ PC	+0.73	
						WJ WA	+1.09	
						WJ WID	+0.95	
				K-5	All students	CTBS RC	+0.42	
						CTBS RV	+0.42	
						CTBS TL	+0.50	
						Gray RC	+0.40	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						Gray RP	+0.46	
						WJ PC	+0.41	
						WJ WA	+0.54	
						WJ WID	+0.59	
						Gray RC	+0.86	
						Gray RP	+0.85	
					Low achiever	WJ PC	+0.49	
						WJ WA	+1.62	
						WJ WID	+1.22	
Chambers et al. (2005)	CQE 1 year	8 schools 577 students	Mostly Hispanic communities in the U.S.	K		WRMT LID	+0.05	+0.30
						WRMT PC	+0.30	
						WRMT WA	+0.27	
						WRMT WID	+0.52	
				1st		WRMT LID	+0.12	
						WRMT PC	+0.29	
						WRMT WA	+0.53	
						WRMT WID	+0.35	
Ross & Casey (1998a)	CQE 2 years	8 schools 356 students	High-poverty schools in Ft. Wayne, IN 75%FRL, 45%minority	K-1	All students	Durrell Oral	+0.21	+0.26
						PC	+0.14	
						Word Attack	+0.46	
						Word ID	+0.22	
					Low achiever	Durrell Oral	+0.31	
						PC	+0.23	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						Word Attack	+0.55	
						Word ID	+0.29	
Datnow et al. (2001)	CQE 4 years	6 schools 398 students	Diverse students in Miami, FL	1-4	ELL	Reading statewide test	+0.61	+0.11
					Non-ELL		+0.04	
Livingston & Flaherty (1997)	CQE 2-4 years	6 schools 828 students	High-poverty multilingual schools in Modesto and Riverside, CA.	K-1	English speaker	Woodcock + DORT	+0.27	+0.40
					Other ESL		0.00	
					Spanish bilingual		+1.39	
					Spanish ESL		+1.32	
				K-2	English speaker		+0.34	
					Other ESL		+0.48	
					Spanish bilingual		+0.42	
					Spanish ESL		+0.68	
				K-3	English speaker		+0.23	
					Other ESL		+0.05	
					Spanish bilingual		+0.03	
					Spanish ESL		-0.09	
Munoz & Dossett (2004)	CQE 3 years	6 schools 349 students	High-poverty schools in Louisville, KY	1-3		CTBS	+0.15	+0.15
Ross et al. (1996a)	CQE 1 year	5 schools 428 students	Tucson, Arizona	1st	All students	Oral Reading	+0.25	+0.41
						PC	+0.33	
						Word Attack	+0.69	
						Word ID	+0.37	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
					Low achiever	Oral Reading	+0.30	
						PC	+0.26	
						Word Attack	+1.10	
						Word ID	+0.36	
					Spanish-Speaking	Oral Reading	-0.07	
						PC	+0.16	
						Word Attack	+0.53	
						Word ID	+0.18	
Ross & Casey (1998b)	CQE 1 year	4 schools 581 students	Suburban schools in Portland, OR. 15% minority	K	All students	Oral Reading	+0.15	+0.06
						PC	+0.01	
						Word Attack	+0.24	
						Word ID	+0.23	
					Low achiever	Oral Reading	+0.26	
						PC	+0.26	
						Word Attack	+0.44	
						Word ID	+0.50	
				1st	All students	Oral Reading	-0.07	
						PC	-0.02	
						Word Attack	+0.10	
						Word ID	-0.10	
					Low achiever	Oral Reading	-0.18	
						PC	-0.25	
						Word Attack	+0.05	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						Word ID	-0.34	
Ross et al. (1994a)	CQE 2 years	4 schools 179 students	African-American students in high-poverty schools in Montgomery, AL	1-2	All students	DORT	+0.52	+0.58
						PC	+0.37	
						WA	+0.80	
						WID	+0.65	
					Low achiever	DORT	+0.66	
						PC	+0.64	
						WA	+1.80	
						WID	+0.94	
Ross et al. (1995)	CQE 3-4 years	4 schools 257 students	Title I schools in Ft. Wayne, IN	K-2	All students	Oral Reading	+0.13	+0.01
						PC	+0.08	
						Word Attack	+0.07	
						Word ID	+0.13	
				K-3		Oral Reading	-0.27	
						PC	-0.19	
						Word Attack	+0.03	
						Word ID	+0.03	
				1-4	All students	Oral Reading	-0.29	
						PC	+0.02	
						Word Attack	+0.18	
						Word ID	+0.08	
					Low achiever	Oral Reading	-0.12	
						PC	+0.33	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						Word Attack	+0.34	
						Word ID	+0.62	
					All students	ISTEP Comp.	+0.49	
						ISTEP Voc.	+0.76	
						Oral Reading	+0.14	
						PC	-0.11	
						Word Attack	+0.17	
				K-1		ISTEP Comp.	+1.11	
						ISTEP Voc.	+1.79	
					Low achiever	Oral Reading	+0.70	
						PC	+0.52	
						Word Attack	+0.69	
Ross et al. (1997)	CQE 2-3 years	4 schools 291 students	A medium-size midwestern city			Word ID	+0.41	+0.17
						ISTEP Comp.	+0.63	
						ISTEP Voc.	-0.15	
						Oral Reading	-0.18	
				K-2		PC	+0.26	
						Word Attack	+0.13	
						Word ID	+0.26	
							+0.28	
						ISTEP Comp.	+0.13	
				1-3		ISTEP Voc.	-0.09	
						Oral Reading	-0.08	

Study	Design Duration	Sample	Sample Description	Grade	Group	Posttest	ES	Study ES
						PC	+0.03	
						Word Attack	-0.03	
						Word ID	+0.17	
Wang & Ross (1999)	CQE 1 year	4 schools 340 students	High poverty African-American schools in Little Rock, AK	1st		DORT	+0.18	+0.08
						WJ PC	+0.10	
						WJ WA	+0.56	
						WJ WID	+0.13	
				2nd		DORT	+0.01	
						WJ PC	-0.09	
						WJ WA	+0.03	
						WJ WID	-0.17	

Note. CR = Cluster randomized; CQE = Cluster quasi-experiment; n = number of effect sizes; ES = effect size; W = White; AA = African-American; FRL = Free/Reduced Lunch; ELL = English Language Learner; H = Hispanic; TOWRE = Test of Word Reading Efficiency; WJ = Woodcock-Johnson; LWID = Letter-Word Identification Test; PC = Passage Comprehension; WA = Word Attack; WRMT = Woodcock Reading Mastery Test; DORT = Durrell Oral Reading Test; CTBS = Comprehensive Test of Basic Skills; RC = Reading Comprehension; RV = Reading Vocabulary; TL = Total Language; WID = Word Identification; Gray = Gray Oral Reading Test; RP = Reading Proficiency; LID = Letter Identification; ISTEP = Indiana Statewide Testing for Educational Progress; Voc. = Vocabulary; Comp. = Comprehension

Appendix 3: Outcomes year by year.

Study	1 year	2 years	3 years	4 years	5 years	6 years
Borman et al. (2007; 2005a; 2005b)	+0.05	+0.18	+0.25			
Chambers et al. (2005)	+0.30					
Nunnery et al. (1996)	+0.19					
Quint et al. (2015; 2014; 2013)	+0.11	+0.14	+0.08			
Ross & Casey (1998b)	+0.06					
Ross et al. (1995; 1994b)	+0.42	+0.46	+0.10	-0.05		
Ross et al. (1996a)	+0.41					
Ross et al. (1996b)	+0.01					
Wang & Ross (1999)	+0.08					
Livingston & Flaherty (1997)		+0.65	+0.40	+0.12		
Ross & Casey (1998a)		+0.26				
Ross et al. (1994a)		+0.58				
Ross et al. (1997)		+0.28	+0.09			
Correnti (2009)			+0.43			
Munoz & Dossett (2004)			+0.15			
Slavin et al. (1993)			+0.59	+0.29	+0.41	+0.43
Datnow et al. (2001)				+0.11		

Mean ES	1 year	2 years	3+ years		
	+0.20	+0.37	+0.14		

Mean effect sizes for each moderator category were calculated by estimated an exploratory model including the same covariates as those shown in Table 1 without an intercept, with the moderator included as a categorical variable and using final and interim report findings.