Review of Educational Research
April 2017, Vol. 87, No. 2, pp. 243–282
DOI: 10.3102/0034654316687036
© 2017 AERA. http://rer.aera.net

Academic Interventions for Elementary and Middle School Students With Low Socioeconomic Status: A Systematic Review and Meta-Analysis

Jens Dietrichson, Martin Bøg, Trine Filges, and Anne-Marie Klint Jørgensen
SFI—The Danish National Centre for Social Research

Socioeconomic status is a major predictor of educational achievement. This systematic review and meta-analysis seeks to identify effective academic interventions for elementary and middle school students with low socioeconomic status. Included studies have used a treatment-control group design, were performed in OECD and EU countries, and measured achievement by standardized tests in mathematics or reading. The analysis included 101 studies performed during 2000 to 2014, 76% of which were randomized controlled trials. The effect sizes (ES) of many interventions indicate that it is possible to substantially improve educational achievement for the target group. Intervention components such as tutoring (ES = 0.36), feedback and progress monitoring (ES = 0.32), and cooperative learning (ES = 0.22) have average ES that are educationally important, statistically significant, and robust. There is also substantial variation in effect sizes, within and between components, which cannot be fully explained by observable study characteristics.

Keywords: systematic review, academic interventions, socioeconomic status, mathematics, reading

Socioeconomic status (SES) is a major predictor of educational achievement (e.g., Björklund & Salvanes, 2011; Currie, 2009; Kim & Quinn, 2013; Sirin, 2005; White, 1982). For example, the results from the Programme for International Student Achievement (PISA) show that the average test score difference between students in the top and bottom 15% of the PISA index of economic, social, and cultural status is about 0.7 to 0.8 standard deviations in the member countries of the Organisation for Economic Co-operation and Development (OECD). This difference is roughly equivalent to 2 years' worth of schooling for that age group. At

The online version of this article has been revised and republished. The original version was published online on January 24, 2017. The revised version was published online on March 12, 2020.

the same time, some low SES students manage to excel in PISA, and the strength of the relationship between SES and test scores differs markedly between countries (OECD, 2010, 2013). These results indicate that overcoming a disadvantaged background is possible, and raise an interesting question: How can interventions propel the achievement of low SES students?

To increase the knowledge about effective interventions, we have performed a systematic review of academic interventions for elementary and middle school students from low SES backgrounds. The review examines interventions implemented by schools, researchers, and local stakeholders, and includes studies that have used a treatment-control design to examine the effects of interventions on standardized test scores in reading and mathematics. The precise research questions guiding the review are the following: (a) What types of interventions can schools and local stakeholders use to increase standardized test scores in reading and mathematics of low SES students in elementary and middle school? (b) What moderates the effect sizes of these interventions? In the next sections, we provide a brief overview of potential reasons why low SES students perform less well academically, explain how interventions may help, survey earlier related reviews, and describe the rationale for the current review.

Socioeconomic Status, Educational Achievement, and How Interventions Might Work

In this section, we first review explanations of the negative correlation between low SES and educational achievement. We then discuss how interventions may help low SES students realize their academic potential by compensating for factors that constrain them.

Explanations of Low SES Students' Educational Achievement

A possible explanation for the achievement differences between high and low SES students is that low SES students have lower innate abilities (see, e.g., Tucker-Drob, Briley, & Harden, 2013, for a discussion). The separation of hereditary factors from environmental factors is inherently difficult, especially when epigenetic effects—that is, heritable genetic changes that are not caused by changes in the DNA sequence but by environmental factors—might be present (e.g., Hackman & Farah, 2009).[1]

However, recent evidence from the United States indicates that hereditary factors are not a major constraint for low SES students (Nisbett et al., 2012). For example, Tucker-Drob, Rhemtulla, Harden, Turkheimer, and Fask (2011) found no significant differences between children in high and low SES families on the Bayley Short Form–Research Edition (see, e.g., Andreassen & Fletcher, 2007)—a test of infant mental ability—at the age of 10 months, but by age 2 children in high SES families scored about one third of a standard deviation higher than children in low SES families. Genes accounted for nearly 50% of the variation in mental ability of high SES children but only a negligible share of low SES children's variation, indicating that the latter are not reaching their full cognitive potential. Rhemtulla and Tucker-Drob (2012) found similar patterns of gene and SES interactions in follow-up tests of mathematics skill at age 4 (but no significant interactions in reading). Fryer and Levitt (2013) found no significant differences on the Bayley Short Form–Research Edition among Hispanic, Asian, Black, and White infants aged 8 to 12 months, although a one standard deviation gap in test scores between Black and White children, which typically differ in SES, has been observed by age 3. In addition, early childhood poverty is a better predictor of later cognitive achievement than poverty in middle or late childhood (Hackman & Farah, 2009). These results are difficult to explain by hereditary factors and suggest that the environment is the main constraining factor for low SES children.

The early childhood environment appears to be one such factor that keeps low SES students from realizing their academic potential. Currie (2009) documented that low SES children have worse health on a very broad range of measures, including fetal conditions, physical health at birth, incidence of chronic conditions, and mental health problems—problems which later influenced educational and labor market outcomes. Family environments are also different between high and low SES children in other aspects thought to affect educational achievement, from early on and onwards. High SES families are more likely to provide a rich language and literacy environment (Hart & Risley, 2003), have different parenting practices, and direct additional resources to early childhood education, health care, nutrition, and enriching spare-time activities (Esping-Andersen et al., 2012). In the United States, low SES children are also less likely to attend center-based care during preschool age (Magnuson & Shager, 2010), and children of low SES parents on average face lower academic expectations from their families (Bradley & Corwyn, 2002; Slates, Alexander, Entwisle, & Olson, 2012).

Low SES students are also likely to live in neighborhoods that are less conducive to educational achievement in terms of, for example, peer support and role models. Also, the skills required to navigate in a disadvantaged neighborhood may be radically different from the skills needed to thrive academically in school. Low SES students may therefore have difficulties decoding appropriate behavior in educational environments (Heller et al., 2015).

The evidence indicates that significant differences in cognitive development and school readiness between high and low SES students are present already prior to school starting age. Do schools contribute to or mitigate this gap? According to Heckman (2006), schools are not a major reason for the inequality in student performance, as gaps in test scores across socioeconomic groups remain stable from third grade onwards. Further evidence is provided by the seasonality in achievement gaps. In the United States, the gap between high and low SES students widen during summer breaks, exactly when schools are out of session (e.g., Alexander, Entwisle, & Olson, 2001; Gershenson, 2013; Kim & Quinn, 2013). Moreover, schools with larger shares of low SES students receive more resources in OECD (2010) countries, which could be an equalizing factor.

The United States is an exception with respect to the student/teacher ratio (OECD, 2010), and teachers have lower expectations of the achievement of low SES students (e.g., Good, Aronson, & Inzlicht, 2003; Timperley & Phillips, 2003). U.S. schools with higher levels of poverty also have lower teacher quality in terms of value added (Glazerman, Protik, Teh, Bruch, & Max, 2013). However, this differential in teacher quality explains only a small share of the test score gap between high and low SES students in some studies (e.g., Chetty, Friedman, & Rockoff, 2014).

On balance, schools do not seem to substantially increase the achievement differences between high and low SES students, but neither do they significantly decrease the gap present at school start (see, e.g., Lipsey et al., 2012, for evidence of the evolution of the achievement gap in terms of the closest equivalent to the outcomes studied in this review, standardized test scores in U.S. elementary and middle schools).

How Interventions Might Work

Our review of the background literature suggests that for reasons that have more to do with nurture than nature, low SES students enter school with fewer of the cognitive and social skills that are closely tied to educational achievement. This evidence points toward a range of interventions that might help low SES students increase educational achievement. Parent training programs, health interventions, role model interventions, or more generally early childhood intervention programs are all types of interventions that may increase academic achievement of low SES children, because they tie into specific domains where low SES children face constraints. Interventions for school-age children—the type of interventions we study—hold particular promise. School interventions can be targeted specifically toward those domains closely related to achievement and where low SES students are at a disadvantage. The literature we discussed above has identified several domains where low SES students are constrained vis-à-vis their high SES counterparts. These domains are likely to correspond to the primary mechanisms of intervention effects: cognitive development, social adjustment (or prosocial behavior), family support, motivational support, increased expectations, and increased pedagogical support (Reynolds, Magnuson, & Ou, 2010).

The broad scope of this review and the ensuing number of possible mechanisms precludes a full discussion of the included interventions' theories of change. But examples of interventions that directly target the domains include tutoring interventions, which provide intensive academic instruction that may compensate low SES children for, among other things, a differential in the access to parental or sibling assistance with school work. Another example is summer reading programs, which addresses the summer dip experienced by low SES children. Similarly, interventions that seek to change mindsets, increase expectations of educational achievement (of students, teachers, and parents), and mitigate stereotype threat address differences in beliefs about, and expectations of low SES students. Other interventions work more indirectly by transferring financial resources or increasing access to higher quality schools. Examples include increased resources to schools, moving low SES students to better quality schools, and programs that incentivize high-quality teachers to teach at low-performing schools.

It is also possible that academic interventions compensate for factors constraining low SES students, even though the interventions do not target the underlying constraining factor. For example, the intensive instruction in a tutoring intervention may compensate a student who is often away from school because of poor health. In this sense, tutoring works as a substitute for an intervention targeting student health directly. Moreover, if differences in educational achievement between high and low SES students stem from differential access to resources in several domains, remedial efforts may need to address more than one domain simultaneously and include a combination of components for greater effectiveness. That is, interventions, or intervention components, may also be complementary. The empirical evidence about which components are complementary and which are substitutes is very scarce in elementary and middle school, however.[2]

Previous Reviews

Given the importance of education for earnings, health, and well-being, finding interventions that effectively improve the educational achievement of disadvantaged children is of considerable importance and a high priority for governments around the world (e.g., UNESCO, 1994). Aspects of interventions targeting educational outcomes for low SES students have accordingly been reviewed previously. Below, we survey the most related and recent reviews of academic interventions that have focused on related outcomes and groups of students that are similar to low SES students.

Four reviews have included target populations close to ours, but focus on a more narrow set of interventions or on different outcomes. Robinson, Ward Schofield, and Steers-Wentzell (2005) examined peer- and cross-age tutoring for minority students using White students as tutors and found positive effects on mathematical, attitudinal, and socioemotional outcome measures for both tutees and tutors. Zief, Lauver, and Maynard (2006) reviewed after-school programs targeting primarily low SES students. They found a small number of studies and weak evidence of positive effects on reading test scores and grade point averages. Kim and Quinn (2013) reviewed summer reading programs and found positive effects on reading test scores for interventions that employed research-based reading instruction and included a majority of low-income children. Wilson, Tanner-Smith, Lipsey, Steinka-Fry, and Morrison (2011) reviewed school completion and dropout prevention programs. A substantial proportion of the targeted students were low SES, and these researchers found large overall positive effects of interventions on dropout and high school graduation rates.

The broad reviews of Slavin and Lake (2008; elementary mathematics programs); Slavin, Lake, and Groff (2009; middle and high school mathematics programs); and Slavin, Lake, Chambers, Cheung, and Davis (2009; reading programs for elementary grades) did not focus on low SES children directly. However, they found no indications that the overall positive effects on test scores differed between low SES students and nondisadvantaged students. The reviews do not contain information about whether the type of programs that show the largest effect sizes—instructional-process programs that, for example, include supplemental tutoring, cooperative learning, classroom management, and motivation interventions—also have the largest effect sizes for disadvantaged students.

Low-achieving students may in many cases overlap with low SES students. Wanzek et al. (2006) reviewed reading programs directed to students in Grades

K to 12 with learning disabilities; and Edmonds et al. (2009); Flynn, Zheng, and Swanson (2012); Wanzek et al. (2013); and Scammaca, Roberts, Vaughn, and Stuebing (2015) reviewed programs for struggling readers in Grades 6 to 12, 5 to 9, above fourth grade, and Grades 4 to 12, respectively. These reviews reported positive effects on test scores in general, but did not report differences over types of interventions in terms of changed instructional methods. Slavin, Lake, Davis, and Madden (2011) also focused on programs directed to struggling readers and did find higher effect sizes, as measured by test scores, for tutoring and classroom instructional process programs, especially cooperative learning. Gersten et al. (2009) examined the effects on test scores of four components of mathematics instruction for students with learning disabilities. They found most support for approaches to instruction (e.g., explicit instruction, use of heuristics), curriculum design, and providing feedback to teachers based on formative assessment data.

The Rationale for the Current Review

The question of what types of interventions or intervention components are most effective for low SES students in elementary and middle school has not been settled in the reviews covered in the previous section. The main objective of our review is to provide policymakers and educational decision-makers at all levels—from governments to teachers—with evidence of the effectiveness of interventions that aim to improve the educational achievement of low SES students. To this end, we have chosen a broad scope in terms of the interventions we include. We have focused the data extraction on the instructional methods of interventions rather than the content taught. Instructional methods are used across subjects and grades, whereas content tends to be more context- and age-specific. Our taxonomy of intervention components contains several components that have not been included in earlier reviews, which either use broader categories of instructional methods, or have examined content.

The scope of our review allows us to compare intervention components in a common and consistent framework. For example, effect sizes in related reviews are often calculated in slightly different ways, using different types of tests. These inconsistencies hinder directly comparing effect sizes among reviews, even for similarly defined intervention components. We have also been able to examine moderators related to implementation, subject, measurement, and the type of participants and tests, as well as dosage and delivery of intervention. Many of these have not been included in meta-regressions in previous reviews using effect sizes based on test scores.

To decide on interventions, educational policymakers should also have knowledge of the costs of interventions; that is, interventions should preferably be cost-effective. The evidence of the cost-effectiveness of interventions in elementary and middle school is very limited, and although our protocol (see the supplementary material, available in the online version of the journal) included an objective to assess also cost-effectiveness, too few studies contained information about costs to make such a comparison meaningful.[3]

Method

This section outlines the methods we have used in the following subsections: inclusion criteria, search strategy, screening and coding, risk of bias, and synthesis procedures and statistical analysis. See our prespecified protocol in Appendix A (available in the online version of the journal) for more information about for example the search strategy and the risk of bias tool.

Inclusion and Exclusion Criteria

Types of Interventions

To be included, interventions had to explicitly aim to improve educational achievement, although they did not have to consist of only academic activities. For example, programs that primarily aimed to reduce criminal behavior or bullying were excluded. We included only interventions that could be implemented by schools or local stakeholders, such as local governments and nongovernmental organizations. Changes to the entire school system, such as changes to the grade system, the national/regional curriculum, and the introduction or expansion of school choice and private schools, were excluded.

We also excluded interventions such as (a) the one described in Fryer (2014) where a bundle of best practices are implemented in low-performing schools, (b) whole-school reform strategy concepts such as Success for All, and (c) studies of different types of schools, such as charter schools. The decision to exclude these interventions was not made because such large-scale interventions are irrelevant to the question of how to improve the educational achievement of low SES students. On the contrary, the available evidence includes many very promising examples of schoolwide reform concepts (e.g., Borman, Hewes, Overman, & Brown, 2003; Epple, Romano, & Zimmer, 2015; Fryer, 2014). However, the nature of such interventions tends to be different than the ones we have included—they are considerably more complex, and contain components not found among the interventions we have included. They may, for example, include a large scale replacement of staff (Borman et al., 2003; Fryer, 2014). The complexity of these interventions sometimes also makes it difficult to identify the intervention components. We therefore deemed these large-scale interventions to be sufficiently dissimilar to the ones we chose to include.

Participants

Included interventions were aimed at low SES students. Most researchers agree on a tripartite nature of SES, which incorporates parental income, parental education, and parental occupation as its three main indicators (Sirin, 2005). Our search included terms aiming to capture all these three aspects, but also some others used in the previous literature to capture populations with low SES. In particular, the share of minority students is often highly correlated with the share of low SES students (Sirin, 2005). To capture studies that lacked information about SES indicators but where a large share of the sample were likely to be low SES students, we interpreted an intervention as being aimed at low SES students if at least 50% of the participants either came from families where parental income, education, or occupation imply low SES status, or from families with minority status. Some studies only reported school or district statistics on SES, instead of individual information on participants. In these cases, we used the aggregate information as an estimate of the share in the sample (see Kim & Quinn, 2013, for a similar procedure).

We included studies of students in elementary and middle school, and excluded studies of interventions in high schools, upper secondary schools, and preschools. The primary reasons for this decision had to do with the comparability of intervention types and the target groups of interventions. Both preschools and high school/upper secondary schools are not compulsory in many countries, unlike elementary and middle school, and they have often curricula that differ more compared with elementary and middle school curricula than these two differ from each other. We do not mean to argue that there are no differences between elementary and middle school, but we deemed the risk of finding noncomparable interventions sufficiently high to exclude interventions in preschool and high school, and sufficiently low to include elementary and middle school in the same analysis. Depending on the country and state, this criterion can include different grades but commonly have included kindergarten up to about Grade 8 to 9. Interventions targeting students receiving special education services within these school settings were included, whereas interventions for students attending special education schools outside a regular school setting were excluded.

Outcome Measures

Included studies used standardized academic tests (e.g., Iowa Test of Basic Skills, Stanford Achievement Test, state- and nationwide tests, norm-referenced commercial tests) to assess the effects of interventions. We restricted attention to standardized tests mainly because earlier reviews of academic interventions indicate that effect sizes tend to be significantly lower for standardized tests compared with researcher-developed tests (e.g., Flynn et al., 2012; Gersten et al., 2009; Scammaca et al., 2015). As researcher-developed tests are usually less comprehensive and more likely to measure aspects of content inherent to treatment, but not control group instruction (Slavin & Madden, 2011), standardized tests may provide a more reliable measure of lasting differences between treatment and control groups. We divided measures into postmeasures and follow-up measures. The former were measured within 3 months after the end of an intervention, and the latter in any period longer than 3 months. If studies collected data twice within the 3-month period, both measures were included.

Study Design

We limited the meta-analysis to primary research with study designs that employed a treatment-control group design. A control group is defined as a nontreatment condition, including waitlist controls. A nontreatment condition does not contain any component of the treatment thought to affect educational achievement. In some of the included studies, students in the control group received what can be described as placebo treatments. We excluded comparison designs, where two groups received alternative treatments, as these effect sizes are not directly comparable to treatment-control designs. Some studies used a staggered introduction of treatment where the control group gets the intervention after, for example, 1 year. The contrast between treatment and control for the following years is then between getting a different length of the intervention. These designs are similar to a comparison design, and we included only the contrasts between treatment and control where the control group had not received the intervention (i.e., the 1 year contrast in the example). Included study designs were randomized controlled trials (RCTs) or quasi-experimental studies (QES). QES included, for example, difference-in-differences designs, matching or statistical controls; that is, QES used some form of nonexperimental technique to mitigate selection bias. Single group pre–post comparisons were excluded.

Settings

Only studies of interventions carried out in the OECD or EU countries were included. This selection decision ensured a certain degree of comparability between settings to align treatment as usual conditions in included studies. Due to language constraints, only studies written in English, German, Danish, Norwegian, and Swedish were included.

Intervention Year

We included studies of interventions performed in or after the year 2000 and excluded interventions performed earlier. If an intervention spanned earlier years in addition to the year 2000, we included the study. Although we wanted the review to be as comprehensive as possible, there is a trade-off between the period span on the one hand and resource use and comparability of interventions on the other (see, e.g., Scammaca et al., 2015, who found smaller effect sizes in more recent studies). We chose the year 2000 as a compromise between these conflicting objectives.

Effect Sizes

To be included in the meta-analysis, a study had to contain information that permitted us to calculate an effect size. We contacted the corresponding author of studies where it was not possible to calculate an effect size, except when the missing information would require running new analyses. The authors were given at least 1 month to reply.

Search Strategy

Relevant studies were identified through electronic searches of bibliographic databases, and government and policy databanks. We searched the following bibliographic databases: Campbell Library, Centre for Reviews and Dissemination Databases, Cochrane Library, Diva-portal.org, EconLit, Education Research Complete, ERIC (Education Resource Information Center), Forskningsdatabasen.dk, Libris, PsycINFO, SocIndex, Social Care Online, and Current Research Information System in Norway.

The search strategy for these databases built on four categories: students, socioeconomic status, outcomes, and study design. For each category, we included a large number of synonyms for the terms in question. The study information had to contain at least one term from each category for the study to be retrieved. We also hand searched the most recent year of the following journals for articles that may not yet have been included in the databases: the *American Educational Research Journal*, the *Journal of Educational Research*, *Learning and Instruction*, and the *Journal of Educational Psychology*.

To mitigate the risk of publication bias, we followed the recommendations in Higgins and Green (2011) and manually searched for unpublished work on websites that publish reports. We used the following websites: OpenGrey (http://www.opengrey.eu/), What Works Clearinghouse (http://www.whatworks.ed.gov), Dansk Clearinghouse for Uddannelsesforskning, (http://edu.au.dk/clearinghouse), European Educational Research Association (http://www.eera-ecer.eu), American Educational Research Association (http://www.aera.net), Deutsche Gesellschaft für Erziehungswissenschaft (http://www.dgfe.de), Skolverket (http://skolporten.com), and forskning.no.

Last, we "snowballed" included articles and the reviews mentioned in the section Previous Reviews by searching the reference lists in these publications to find studies that our electronic database search might have missed. We also used the following, less closely related, reviews to snowball references (topic and target population in parentheses): Ritter, Albin, Barnett, Blankenship, and Denny (2006; volunteer tutoring for general student populations); Goodwin and Ahn (2010; morphological interventions for children with literacy difficulties); Alfieri, Brooks, Aldrich, and Tenenbaum (2011; discovery-based instruction for general student populations); Dexter and Hughes (2011; graphic organizers for students with learning disabilities); Cheung and Slavin (2012; technology applications for general student populations); Kyndt et al. (2013; cooperative learning for general student populations); de Boer, Donker, and van der Werf (2014; attributes of interventions for general student populations); and Reljic, Ferring, and Martin (2015; bilingual programs to European students).

Screening and Coding Studies

Screening was done by four review team assistants together with the study authors. First-level screening used abstracts and titles to exclude clearly irrelevant studies. We piloted the criteria with all four assistants until a higher than 85% agreement with the study authors was achieved. Remaining disagreements involved, to a very large extent, assistants including more studies than study authors. Studies that were included in the first level were obtained in full text and screened by the same team. We did not double-screen all studies in the second-level screening. However, research assistants were instructed to only exclude studies when they were certain of a study not meeting the inclusion criteria, and to discuss all uncertain cases with a study author. A consensus decision was reached in all uncertain cases.

The numerical coding was done by at least two of the study authors. For each study, the coding of intervention components was discussed between at least three members of the team, at least two of whom were study authors. Coding of other moderator variables was performed by at least two research assistants per study. Coding was not duplicated and blind. Instead, to minimize errors, one coder coded a full study, and thereafter a second coder first read the study and then critically assessed the coded information. Discrepancies between coders were resolved by discussion. Research assistants were graduate students in sociology or economics, and the study authors involved in coding (the first three authors) all had doctoral degrees.

Risk of Bias

The purpose of assessing the risk of bias of each included study is to ascertain whether the study methodology is sufficiently rigorous such that the effect estimate is more likely to inform than mislead a meta-analysis. That is, we assessed internal validity. We did this using an extended version of the Cochrane risk of bias tool (Higgins & Green, 2011). The extension permits the assessment of studies that use a nonrandomized study design. Studies are assessed for bias in the following domains: selection and confounding bias, performance bias, detection bias, attrition bias, outcome reporting bias, and other bias. A study is scored per outcome in each domain according to whether there is a too high risk of bias or not. Studies that are scored as having a high risk of bias in one or more domains were excluded from the meta-analysis.

Each study was assessed by at least one of the authors. In cases where the assessment would lead to exclusion of the study from meta-analysis, a second author independently assessed the study, and any disagreement in the assessment was resolved through discussion.

Coding of Intervention Components

To characterize the instructional method used in the intervention, we coded whether the interventions contained one or more of the major components listed below. These components were not prespecified, but developed and adapted during the coding process. We partly used previous reviews and author-reported classifications in included studies, and partly used an iterative process to construct component categories.

After School Programs

This category contains interventions that are implemented in after-school settings. Examples include local initiatives where schools, municipalities, community organizations, and firms work together to provide academic support and a safe local environment.

Coaching/Mentoring of Students

We have separated interventions in this category from tutoring interventions (see below), as coaching and mentoring do not have to involve any pedagogical support. Activities can be to help students with the choice of courses to take, and provide role models, for example.

Coaching/Mentoring of Personnel

This category includes programs that provided teachers or other school personnel with coaches or mentors. We separated this component from personnel development interventions that seek to develop more general skills with regard to teaching or management; the coaching and mentoring in this category are often connected to the implementation of a specific reading or mathematics program.

Computer-Assisted Instruction

Computer-assisted instruction interventions used computers and software programs to try to enhance student achievement. Examples include programs that provided students and teachers direct diagnostic feedback about students' reading progress, and programs that helped teachers implement supplemental mathematics instruction.

Content Changes

This broad category consisted of interventions that *only* changed the content taught and not the methods of instruction. The category included changes to the whole curriculum, but most studies in the meta-analysis examined material that supplemented a broader curriculum. Examples included teaching content in English and students' native language instead of just in English, and increased focus on phonemic awareness, vocabulary, reading comprehension, the natural numbers and operators, geometry, and measurement (although there is only one study of mathematics in this category).

Cooperative Learning

Cooperative learning, or peer-assisted learning, referred to interventions where students work together in pairs or small groups in a systematic and structured manner. Examples included students acting as pedagogical instructors for each other, as when more able students help less able students.

Feedback and Progress Monitoring

This category included interventions that added a specific feedback or progress monitoring component, where teachers or students received detailed information about the students' development. The objective was often to customize instruction to the individual student's needs. Note that tutoring and cooperative learning are also likely to contain increased feedback, but because such feedback is embedded in the regular set up of these programs, these interventions are not coded in this category. Interventions had to add an extra component of feedback or progress monitoring to be coded here.

Incentives

Incentive programs intended to increase the academic performance of students were included in this category. The incentives needed not be monetary, and could target students as well as teachers and families. Examples included interventions where students were paid to read books, families got extra resources in relation to students' academic performance, or teachers were given bonuses for teaching in low-performing schools.

Increased Resources

Interventions in this category simply increased resources without entailing a specific change to the pedagogical content or methods. Only two studies of this type were included in the meta-analysis: one increased resources by increasing the supply of teachers, and the other gave stipends to students to attend higher performing private schools.

Personnel Development

Personnel development implied that school personnel, mainly teachers or principals, received training or further education. Training provided just in order for teachers to implement a specific program, intervention, or test were not coded in this category. Included personnel development interventions had a broader objective.

Psychological/Behavioral Interventions

Psychological/behavioral interventions focused on improving educational achievement through improving social–cognitive skills, mitigating problematic behavior, and changing expectations or beliefs. Examples included prevention programs for acting out students, school-wide programs in socioemotional learning, play therapy, and stereotype threat interventions.

Small-Group Instruction

Interventions in this category included instruction where students are placed in groups smaller than regular class sizes. These interventions differed from those in which learning in small groups are built in, such as cooperative learning and tutoring. There was no cooperative learning element explicitly included in the interventions coded in this category, and the groups were larger than what normally counts as tutoring (here defined as more than five students per group).

Summer Programs

These interventions were administered during a summer break. A typical intervention supplied books to students to read and work with during the summer; the book selection was done in cooperation with librarians or reading coaches. Most programs used some structured introduction and evaluation before, during, and after the summer break, for example, by sending out report cards or phone calls from teachers during the summer.

Tutoring

Tutoring interventions were activities where students got supplemental pedagogical support from an instructor, either one-to-one or in a small group (five students or fewer). Tutors could be volunteers, paid non-teachers, or professional teachers. The interventions included in the tutoring category were often highly structured programs (e.g., manual based) implemented over a limited time period, typically 12 to 20 weeks.

Statistical Method

Effect Size Metrics

For continuous data, standardized mean differences (SMDs) were calculated. We used covariate adjusted mean differences whenever available, and the unadjusted standard deviation to calculate SMDs. We used Hedges' *g* to get an unbiased estimate of the effect size in small samples. Hedges' *g* and its standard error are calculated as follows (Lipsey & Wilson, 2001):

$$g = \left[1-\left(3/\left(4N-9\right)\right)\right]\left[\left(X_1 - X_2\right)/s_p\right] \tag{1}$$

$$SE_g = \left[\left(N/n_1 n_2\right)+\left(g^2/2N\right)\right]^{1/2} \tag{2}$$

where $N = n_1 + n_2$ is the total sample size, X_1 and X_2 is the mean in each group, and s_p is the pooled standard deviation defined as

$$s_p = \left[\left((n_1-1)(s_1)^2+(n_2-1)(s_2)^2\right)/(n_1+n_2-2)\right]^{1/2} \tag{3}$$

Here, s_1 and s_2 denote the unadjusted standard deviation of the treatment and control group. We used the standard deviations from posttests because some RCTs did not include a pretest. For the same reason, we have consistently used posttest mean differences rather than gain or change scores.

Two studies reported only dichotomous outcomes. We transformed these effect sizes to SMDs by using the methods in Sánchez-Meca, Marín-Martínez, and Chacón-Moscoso (2003). One study reported SMDs at the school-level, whereas all other studies reported student-level effect sizes. We obtained information from the study authors to convert the effect size to the student level.

Nine studies reported a different effect size where the mean difference was standardized using the control group's standard deviation (i.e., Glass's Δ). We included these effect sizes, and used the same small sample adjustment as in Equation (2). Furthermore, 10 studies used the district-wide standard deviation, but did not include information about the respective standard deviation for treatment and control group. We included these effect sizes under the assumption that the standard deviation in treatment and control groups are identical to the overall level. Judging from studies that standardize in this way and report the standard deviation of treatment and control groups separately, this decision probably results in an overestimation of the pooled treatment and control group standard deviation. A similar problem arises when studies are based on population samples at different levels. For example, the variance in a study conducted in several school districts includes a district-level variance component, which a study of schools in one district does not. Studies with more levels may therefore yield smaller effect sizes for the same magnitude of intervention effects, simply because the raw standard deviation is larger (Lipsey et al., 2012). We examined whether these measurement issues influenced effect sizes in the meta-regressions (described below). Where studies had missing summary data, such as missing standard deviations, we calculated these where possible from, for instance, *F* ratios, *t* values, and chi-squared values using the methods in Lipsey and Wilson (2001).

Data Synthesis

In subgroup analyses and meta-regressions, we used random effects models, weighting by the inverse variance of effect sizes. Studies with more participants were therefore given more weight, all else being equal. Compared with fixed effects models, studies with small samples receive a higher weight in random effects models because they provide information about the variation around the mean effect size. Given the broad scope of the review, we believed that a random effects model was more appropriate. In subgroup analyses, we also provide a graphical display (forest plot) of effect sizes, and report 95% confidence intervals.

To identify the characteristics of study methods, interventions, and participants that were associated with smaller and larger effects on the various outcomes, we used mixed-model meta-regressions (see, e.g., Lipsey & Wilson, 2001). We examined several moderators. Gender is measured as the share of girls. If available, we have the proportion in the treatment group. Otherwise, we approximated this proportion by the proportion in the whole sample or the school using school district information.

Studies often reported either the mean age or the grade level, and in addition, many interventions spanned several grades and ages. We therefore included an indicator variable separating two broad grade levels, corresponding approximately to elementary and middle school. The indicator for intervention in elementary school took the value 1 if students were in Grades K to 5 or had mean ages up to 11, and 0 otherwise. If a study spanned both elementary and middle school, we coded it in the category where most students or grades belong.

We also coded four types of proxies for SES: income (typically measured by free/reduced lunch status), parental education, parental occupation, and minority status. The first and the last types were by far the most commonly used, and we included the share of low income and minority students as moderators. However, a relatively large number of studies did not provide information about both these variables, and they were highly correlated. In some regressions, we therefore included an indicator for whether the proportion of students was more than 75% on at least one of these two measures. If a study had only reported a range or a lower bound of, for example, low income students, the midpoint of that range or the lower bound had to be at least 75% for the indicator variable to take the value 1. We used the midpoint and the lower bound for the continuous variables in a similar fashion.

Regarding treatment modality, we coded whether the intervention was provided by professionals, received individually or in groups, and whether the providers received training. Interventions could be different in terms of dosage: they were implemented over varying lengths of time, and included different number of sessions and hours per week. Duration was measured in weeks, and we translated a school year into 40 weeks, or approximately 9 months. If ranges were reported, we used the midpoint of the range for all three variables. If the actual number of hours or sessions was not available, we coded the intended or recommended number. Not all the three variables were relevant for all types of interventions. It was, for example, difficult to determine the frequency of incentive programs. Therefore, we did not use the three variables describing the dosage of an intervention in all analyses.

We included an indicator variable for whether the implementation quality or treatment fidelity was assessed during the intervention. We interpreted the lack of information about an assessment as no assessment taking place. We also included an indicator for RCTs and an indicator for mathematics tests.

Last, we included three moderators that related to the measurement of effects. Glass's Δ indicated whether the SMD was standardized with the control group's standard deviation. We had an indicator for whether the SMD had been standardized with a standard deviation from a super-population (e.g., district or state) instead of the pooled standard deviation of the treatment and control group, or if the number of included schools is larger than nine, or the number of districts is more than one (the median over studies in both cases). For the third measurement variable, we coded whether the particular assessment was a test of a subdomain of reading or mathematics (e.g., decoding or number sense) or a more general test. Tests of two or more domains were coded as general. Subtests may be more likely to be inherent to treatment, but it seemed reasonable that interventions that target subdomains of reading and math be tested on whether they affected these subdomains. Either way, study differences in the use of subtests might be an important and novel explanation for variation in effect sizes.

Dependent Effect Sizes

Studies with multiple intervention groups with different individuals and studies using multiple tests for the same intervention groups were included in this review. To avoid problems with dependence between effect sizes, we used robust variance estimation when we conducted meta-regression (Hedges, Tipton, & Johnson, 2010).[4] Simulation studies show that this method needs around 20 to 40 studies included in the data synthesis in order to produce reliable standard errors (Hedges et al., 2010; Tipton, 2015). In subgroup analyses, we used a synthetic effect size (the average) in order to avoid dependence between effect sizes. This method provides an unbiased estimate of the mean effect size parameter but overestimates the standard error and tests of heterogeneity are rejected less often (Hedges, 2006).

Missing Moderator Data

Data were missing on some study characteristics. In order to examine these moderators without biasing the estimates by leaving out these effect sizes, we imputed values using multiple imputation techniques common to meta-analyses.[5] To produce valid estimates, multiple imputation rests on the assumption that data is missing at random; that is, missing observations do not depend on unobserved data, conditional on the observed data. This assumption is not possible to test directly, which calls for caution in the interpretation of variables with missing data.

Heterogeneity

Heterogeneity was assessed with the chi-squared (Q), the I^2, and τ^2 statistics (Higgins, Thompson, Deeks, & Altman 2003). The statistics provide an answer to whether the dispersion of effect sizes is likely to be due to sampling error (Lipsey & Wilson, 2001). The I^2 statistic is independent of the number of studies, which the Q test is not, but it is not independent of the size of studies (Rücker, Schwarzer, Carpenter, & Schumacher, 2008). The τ^2 is independent of both the number and size of studies, but dependent on the scale of the effect size (Borenstein, Hedges, Higgins, & Rothstein, 2009). Therefore, we focused on the I^2 and the τ^2 statistics, and commented on the Q test when it added information. When there were enough studies to yield a reasonable estimate and the mean effect size was of interest, we used the τ^2 statistic to calculate the prediction interval (Borenstein et al., 2009). This interval addresses the dispersion of effect sizes, rather than the accuracy of the mean effect size (as a confidence interval does). We used a 95% prediction interval, which indicates the range in which we should expect effect sizes from a new study to end up 95% of the time.

Clustered Assignment of Treatment

In cluster RCTs, participants are randomized to treatment and control groups in clusters. QES may also include clustered assignment of treatment. In such studies, standard errors may be biased if the unit-of-analysis is the individual and no adjustment is made for the clustering (Higgins & Green, 2011). In most studies using a clustered assignment of treatment, we lack information to properly correct individual effect sizes and standard errors. Thus, we adjusted effect sizes using estimates of the intracluster correlation, and assumed equal cluster sizes if no information is provided in the study. The majority of studies in the meta-analysis are multisite studies, and we used the formulas in Hedges (2007) to correct effect sizes and standard errors. To calculate an average cluster size, we divided the total sample size by the number of clusters (typically the number of classrooms or schools). The intra-cluster correlation differs by grade and subject, and we used the estimates (without pretest adjustment) for schools with low SES reported in Table 4 in Hedges and Hedberg (2007), averaged over Grades K to 2, 3 to 5, and 6 to 9, respectively. Because we could not convert each effect size individually, this analysis may be biased. We therefore only adjusted for clustered assignment in the sensitivity analysis, and not in the main analyses.

Results

We first present the outcome of the search and screening process. In the second section, we display overall effect sizes for the studies included in the meta-analysis, divided by study design, subject, and the timing of outcome measurement. The third section describes results from meta-regressions using study characteristics as moderators. The fourth examines the effect sizes of intervention components. Last, we present the results from sensitivity analyses.

Search and Screening Results

We ran the searches of electronic databases during October 2014. The total number of potentially relevant studies constituted 10,766 unique studies. A total of 1,432 potentially relevant studies were found through other sources. Figure 1 illustrates the literature search and screening. After the first level screening on titles and abstracts of studies from the electronic databases, in total 1,137 references were retrieved for full text screening. Of these, 193 met the inclusion criteria. Forty-four of these were identified from snowballing reference lists. No studies from the hand search or from the search of the grey literature were included. A total of 101 studies were included in the meta-analysis (see Appendix B, available in the online version of the journal).

The difference between the number of studies that met the inclusion criteria and the number included in the meta-analysis depends on the following: first, some studies were excluded due to a high risk of bias (70 studies in total). Of studies deemed to have too high risk of bias, 23 were QES that lacked information about pretest score differences for treatment and control groups. Other categories consisted of QES that did not control for other potential confounders, had noncomparable treatment and control groups (e.g., used students who accept and refuse treatment, respectively), or confounded treatment effects with school (or class) effects by, for example, assigning treatment at the school level

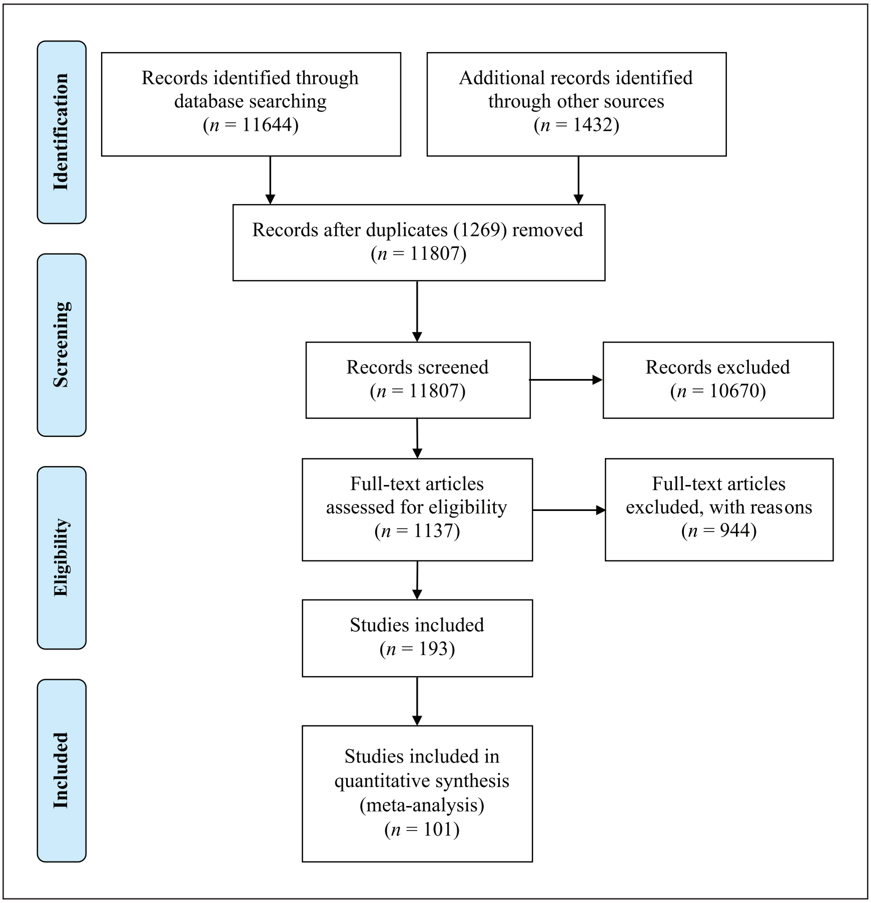
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 11644)
Additional records identified through other sources (n = 1432)
Records after duplicates (1269) removed (n = 11807)
Records screened (n = 11807)
Records excluded (n = 10670)
Full-text articles assessed for eligibility (n = 1137)
Full-text articles excluded, with reasons (n = 944)
Studies included (n = 193)
Studies included in quantitative synthesis (meta-analysis) (n = 101)

FIGURE 1. *Flowchart of the search and screening process.*

and having only one treated and one control school. Many of these studies had two or more of these problems and 32 studies were excluded for one or more of these reasons. The remaining 15 studies were deemed to have too high risk of bias for more idiosyncratic reasons.

Twenty studies were excluded from the meta-analysis because they did not contain information to calculate an effect size or its standard deviation, and we were unable to obtain the information from study authors. Last, two studies were excluded because they used identical samples to already included studies. In the first case, we included the study that had been published, and in the second case, one report contained information about more treatments than the other, and we chose the one with more information.

Table 1 displays descriptive statistics. Ninety-five percent of the studies were from the United States. There was one study each from Australia, Canada, Chile, Germany, and Sweden. The number of students ranged from 19 to 548,903, and 76% of studies were RCTs. The sample was skewed toward reading, but 33 studies contained at least one test in mathematics (many studies tested outcomes in both math and reading). It is notable that some studies have missing information on study characteristics and that the distributions of some of these characteristics are skewed. For example, 78% of studies have been performed in elementary school and 82% include some form of training for providers. We describe the intervention components in terms of, for example, number of studies in the section, Effect Sizes by Intervention Components.

Effect Sizes by Subject, Study Design, and Measurement Timing

In this section, we calculated weighted average effect sizes, separating between studies that have used standardized tests in reading (including spelling tests) and mathematics, between RCTs and QES, and between end-of-intervention and follow-up measurements. The weighted average effect size and the individual study effect sizes (averaged over treatments and tests if applicable) are displayed graphically in forest plots. The forest plots show the average effect size of each study (the center of each grey square) and its 95% confidence interval. The black bars indicate the length of the interval. The size of the grey square is proportional to the weight a study contributes to the overall average effect size. The diamond at the bottom of each forest plot displays the weighted average overall effect size and its confidence interval. The midpoint of the diamond indicates the magnitude of the effect size. If the diamond does not cross the zero line, then the overall effect size is statistically significant ($p < .05$).

Reading

Figure 2 shows the effect sizes of the 66 RCTs that have used standardized tests of reading at the end of intervention. The overall effect size is 0.09, which is statistically significant (95% confidence interval [CI] [0.06, 0.13]). There is considerable heterogeneity between effect sizes, which range from −0.23 to 0.99. The I^2 statistic is 75.2%, and the τ^2 is 0.007, yielding a 95% prediction interval ranging from −0.076 to 0.26. Both statistics imply that there is systematic variation between effect sizes.

Figure 3 shows the corresponding forest plot for the 21 QES. The overall weighted effect size is 0.18, which is statistically significant (95% CI [0.14, 0.31]). The range of effect sizes (−0.13 to 0.82) and the tests of homogeneity indicate that the variation among the effect sizes is unlikely to be due to sampling error alone. The I^2 statistic is as high as 94.4% and τ^2 is 0.0212. The 95% prediction interval ranges from −0.072 to 0.52.

Only a few RCTs (5) and QES (2) tested follow-up reading outcomes, and we have therefore omitted these forest plots. The weighted averages are 0.22 and −0.03, respectively, for RCTs and QES. Given the small number of studies with each study design, we did not assess heterogeneity of effect sizes.

TABLE 1

Descriptive statistics for study context, design, outcome assessment, participant, and intervention delivery characteristics

Study characteristics	*k*	*M*	*SD*	*Mdn*	Range
Study context					
% Performed in United States	101	95	22	1	0–1
Participants[a]	101	10617.8	63341.6	243.0	19–548,903
Number of schools[a]	89	56.8	324.3	9	1–3,053
Number of districts[a]	81	8.2	4.6	1	1–415
Study design characteristics					
% RCT	101	76	42	1	0–1
% Implementation fidelity assessed	101	74	44	1	0–1
Outcome assessment					
% Assessed on subdomain[b]	101	55	48	1	0–1
% Standardized with large *SD*[c]	101	35	47	0	0–1
% Glass's Δ effect size	101	9	29	0	0–1
Participant characteristics					
% Girls	82	46.6	8.3	48.0	16–73
% Elementary school	101	78	48	1	0–1
% More than 75% low SES	101	62	48	1	0–1
% Minority	95	73.3	24.2	78	1–100
% Low income	87	72	18.5	74	22–100
Intervention characteristics					
% Math	101	21	33	0	0–1
Duration in weeks	89	30.3	27.2	20.0	1–160
Number of sessions	53	54.2	36.3	48.0	6–200
Hours per week	58	2.6	2.6	1.8	0–30
% Individual recipient	95	0.41	49	0	0–1
% Delivered by professionals	87	60	49	1	0–1
% Training provided	83	82	39	1	0–1

Note. RCT = randomized controlled trial; SES = socioeconomic status; *k* = number of unique study samples; *M* = mean; *SD* = standard deviation. All variables are measured at the study level.

[a] The number of participants, schools, and districts refer to the total sample of treatment and control groups.

[b] Indicates whether the outcome is based on a test of some subdomain of reading or mathematics (e.g., decoding or number sense), or a more general test. Tests of two or more domains have been coded as general.

[c] Indicates that the outcome has been standardized with a standard deviation from a sample population that is broader than usual, defined as either standardizing with a super-population (e.g., district or state) instead of the pooled standard deviation of the treatment and control group, or if the number of included schools is larger than nine, or the number of districts is more than one (the median over studies in both cases).

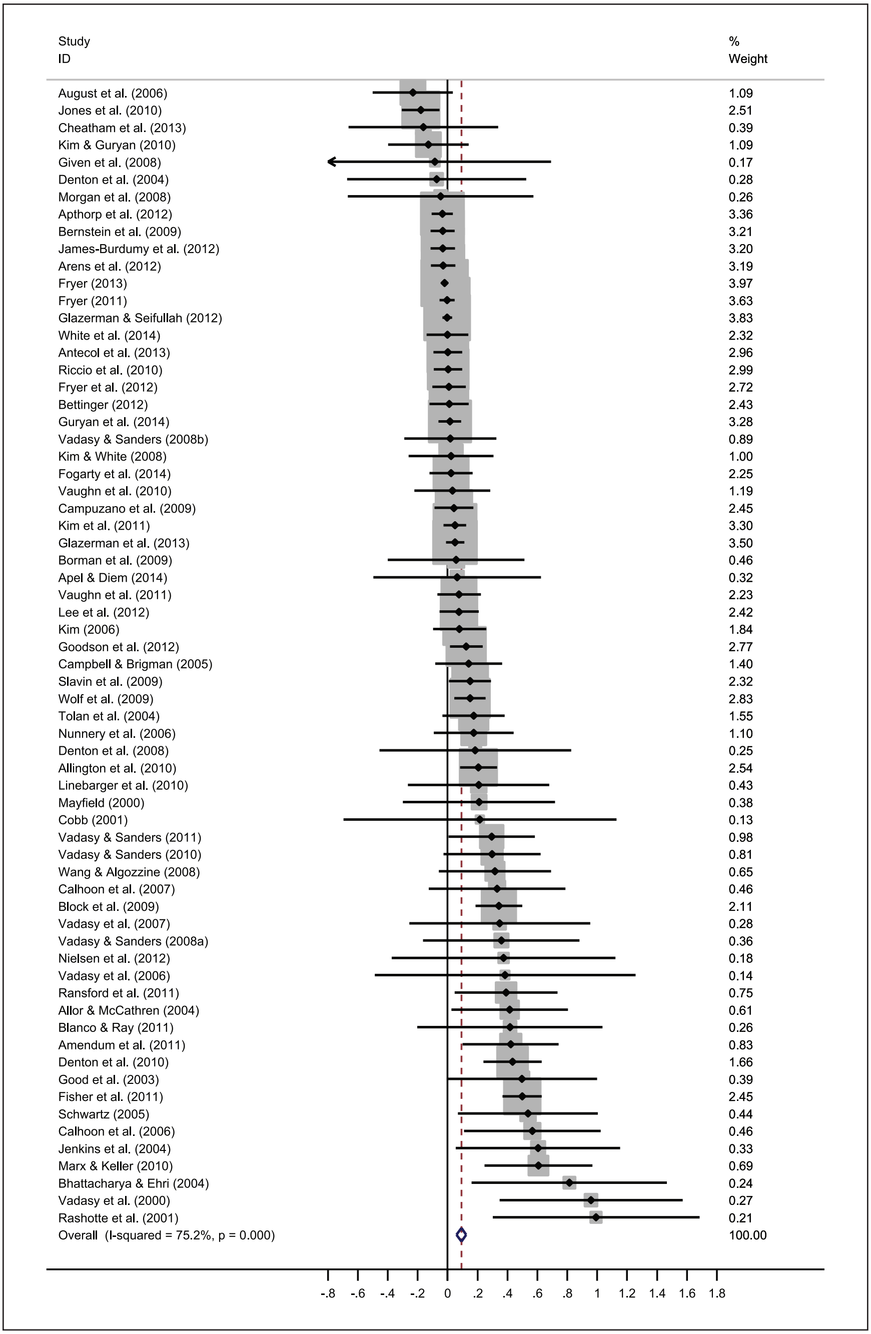

FIGURE 2. *Weighted effect sizes in randomized controlled trials of reading interventions with outcomes measured at the end of interventions.*

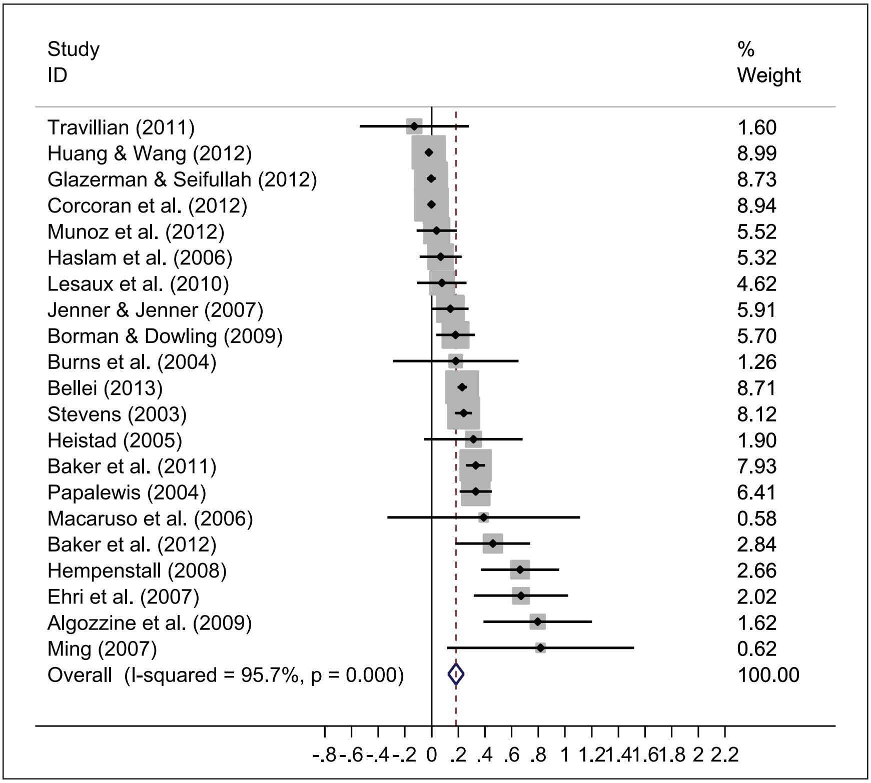

FIGURE 3. *Weighted effect sizes in quasi-experimental studies of reading interventions with outcomes measured at the end of interventions.*

Mathematics

Figure 4 shows the effect sizes from the 25 RCTs that have used standardized tests of mathematics at the end of intervention. The overall effect size 0.10 is similar to that of reading and statistically significant (95% CI [0.05, 0.14]). The effect sizes range from −0.12 to 1.27. The I^2 and τ^2 statistics are again high, 85.5% and 0.008. The 95% prediction interval ranges from −0.08 to 0.28. Eight QES used standardized tests of mathematics at the end of interventions. The weighted average is smaller than the one obtained for RCTs, 0.05, and not statistically significant. No included study has tested a follow-up outcome in math with an RCT, and only two QES has done so. The weighted average effect size for the latter category is 0.05 (95% CI [0.02, 0.08]). Due to the low number of studies, we did not show these forest plots.

In sum, the results indicate positive and mostly significant weighted average effect sizes. There is considerable heterogeneity within the categories created by subject, study design, and measurement timing, which motivates the analyses in

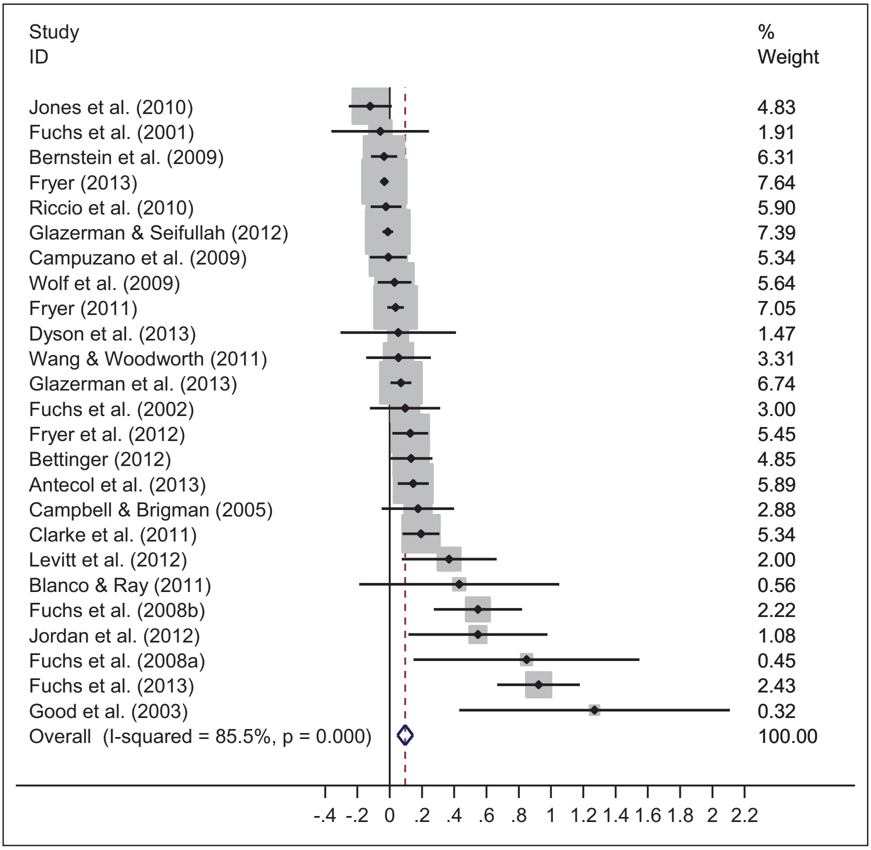

FIGURE 4. *Weighted effect sizes in RCTs of math interventions with outcomes measured at the end of interventions.*

the next section where we report results from meta-regressions that examine the associations between effect sizes and moderators.

Meta-Regressions

Three studies included only follow-up measures, and the total number of studies that included such measures is also small. Because follow-up measures are not fully comparable to end-of-intervention measures, we do not include follow-up measures in the analyses below. We pool all remaining studies in the regressions and have included indicators for study design and subject,[6] since the previous section revealed relatively small differences between the categories created by dividing studies by subject and study design, with the partial exception of reading outcomes between RCTs and QES.

Table 2 displays the results of four meta-regressions, which include 446 effect sizes from 98 studies. The first two columns contain only moderators without

TABLE 2

Results from meta-regressions examining differences in effect sizes attributable to moderators

Moderator	(1)	(2)	(3)	(4)
Study design characteristics				
RCT	–0.0606	–0.101*	–0.097	–0.0916
	(0.0428)	(0.0461)	(0.0536)	(0.0523)
Implementation assessment	0.0025	0.0101	0.0052	0.0141
	(0.0401)	(0.0423)	(0.0517)	(0.0494)
Outcome assessment				
Sub-domain	0.170**	0.124*	0.125*	0.130*
	(0.0411)	(0.0490)	(0.0531)	(0.0539)
Standardised with super-Population	–0.0393	–0.0301	–0.0393	–0.0458
	(0.378)	(0.0416)	(0.0513)	(0.0501)
Glass's Δ	–0.124*	–0.0721	–0.0847	–0.103*
	(0.0432)	(0.0501)	(0.0587)	(0.0483)
Participant characteristics				
Share of girls			0.00397	0.0039
			(0.00341)	(0.0036)
Elementary school		0.0601	0.0718	0.0836
		(0.0449)	(0.0465)	(0.0424)
At least 75% low SES		–0.0301	–0.0338	
		(0.0414)	(0.0485)	
Share of minority				–0.0022*
				(0.0010)
Share of low income				0.0019
				(0.0015)
Intervention characteristics				
Math		0.0371	0.0376	0.0335
		(0.0405)	(0.0411)	(0.0386)
Duration in weeks			–0.00004	0.0000
			(0.00083)	(0.0008)
Individual recipient			–0.0304	–0.0299
			(0.0500)	(0.0458)
Delivered by professionals			0.0433	0.0346
			(0.0491)	(0.0505)
Training provided			–0.0200	
			(0.0508)	
Tutoring component		0.159*	0.194**	0.167*
		(0.0679)	(0.0679)	(0.0702)
Constant	0.161**	0.131	0.118	0.0751
	(0.0542)	(0.0766)	(0.103)	(0.0729)
Multiple Imputation	No	No	Yes	Yes
Effect sizes (n)	446	446	446	446
Number of studies (k)	98	98	98	98
I^2	87.0	86.4	84.2	83.8

Note. RCT = randomized controlled trial; SES = socioeconomic status. Robust standard errors in parentheses.

**p* < 0.05. ***p* < 0.01.

missing values. In the last two columns, we have imputed missing values for moderators using multiple imputation techniques. In Column (1) and all other columns, we have included indicator variables that describe the type of study design and the measurement of effect sizes. In Column (2), we have added all other moderators without missing observations.[7] Column (3) also includes variables with imputed values for missing observations.[8] Our low SES indicator, which is included in Columns (1) to (3) is a catch-all that does not distinguish between the multiple dimensions of low SES status, and may confound minority status with low SES. Column (4) addresses this issue by controlling for the share of minority participants and the share of low-income participants.

There are five variables, which are sizeable in comparison to the overall weighted average effect size of 0.13. The indicator for study design (RCT) is always negative, with a magnitude around −0.10, and significant in one specification (see Table 2). Tests conducted on subdomains are positive and large, around 0.13 in most specifications, and significant in all specifications. Reported effect sizes that use Glass's Δ are about 0.10 smaller and significant in specifications (1) and (4). Effect sizes from interventions including a tutoring component are at least .16 larger and significantly so. The association between effect sizes and the share of minority students is negative, small and significant, whereas the share of low income participants is positive, small, and not significant. Note though that these variables are highly correlated (0.65).[9] The estimates of most other moderators are relatively small in magnitude and are not precisely estimated. An exception is the indicator for interventions conducted in elementary school. The estimate is around .06 to .08.

The coefficients of moderators without missing observations are relatively stable across specifications, and the results are not particularly sensitive to outliers. Removing the 5% smallest and the 5% largest effect sizes yield very similar estimates (results available on request). A large share of the variation between effect sizes is left unexplained in all specifications; the I^2 statistic varies between 84 and 87%. As indicated above, a remaining candidate for the dispersion in effect sizes is the intervention components. We examine the average effect size by component next.

Effect Sizes by Intervention Components

In Table 3, we report the number of studies using each component, the weighted average effect size per component, its 95% confidence interval, and the I^2 and τ^2 statistics. Figure 5 shows weighted average effects sizes and confidence intervals graphically. Only effects at the end of intervention are included.[10] There are three studies that contain more than one intervention and where the interventions consist of different components. In these cases, the intervention containing, say, computer-assisted instruction contributes to the *computer-assisted instruction* component, and the treatment containing content changes contributes to the *content changes* component. The meta-regressions suggest that we can pool math and reading interventions in the same analysis, but RCTs had lower effect sizes on average compared with QES. Because there are few studies for most components, we conduct a pooled analysis here. We examine the sensitivity of the results in the section, Sensitivity Analysis.

TABLE 3

Number of studies, weighted average effect size, confidence interval, I^2 and τ^2 for each intervention component

Intervention component	*k*	Average effect size	95% Confidence interval	I^2	τ^2
Incentives	8	0.01	[–0.02, 0.04]	59.86	0.00
After school programs	3	0.02	[–0.06, 0.11]	36.91	0.00
Summer programs	8	0.03	[–0.06, 0.12]	47.68	0.01
Coaching students	11	0.04	[–0.14, 0.22]	60.40	0.03
Psychological interventions	7	0.05	[–0.16, 0.26]	73.33	0.05
Personnel development	8	0.07	[–0.05, 0.18]	97.16	0.02
Increased resources	2	0.08	[0.01, 0.15]	0.00	0.00
Computer-assisted instruction	9	0.11	[–0.01, 0.22]	69.52	0.02
Coaching personnel	10	0.16	[0.04, 0.28]	94.21	0.03
Content changes	9	0.19	[0.06, 0.31]	76.83	0.02
Cooperative learning	10	0.22	[0.10, 0.34]	74.05	0.02
Small-group instruction	4	0.24	[0.00, 0.48]	86.13	0.05
Feedback and progress monitoring	5	0.32	[0.18, 0.47]	70.06	0.02
Tutoring	36	0.36	[0.26, 0.45]	65.31	0.05

Note. k = number of unique study samples. Effect sizes are averaged over studies and intervention component, using a random effects model and inverse variance weights to produce the weighted average effect size.

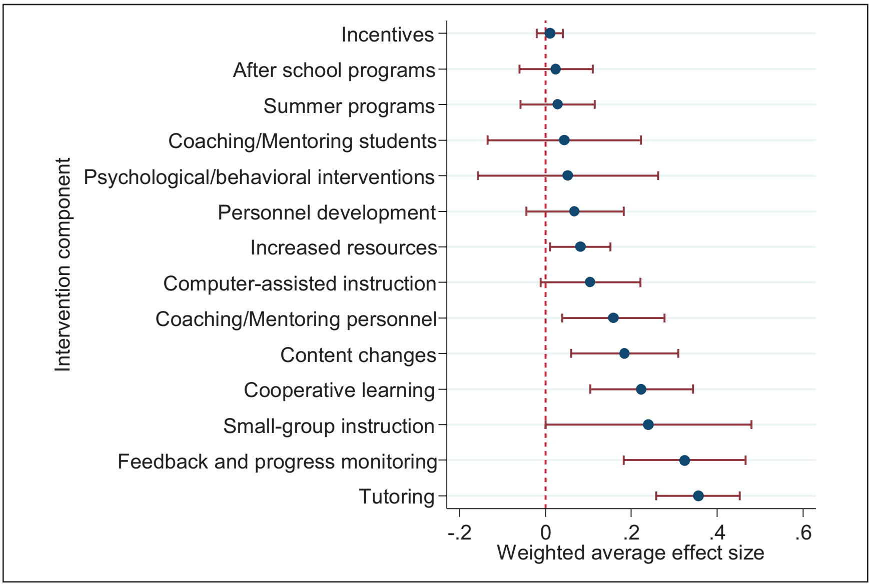

FIGURE 5. *Weighted average effect sizes by component*

All components have positive weighted average effect sizes, but many are small. *Incentives, after school programs, summer programs, coaching/mentoring students*, and *psychological/behavioral interventions* all have weighted effect sizes below, or just above 0.05, none of which are statistically significant. Some components, in particular psychological/behavioral interventions and coaching/mentoring students, have rather wide confidence intervals. A wide interval may indicate, among other things, that there are both interventions with large positive effects and large negative effects, which is the case for psychological/behavioral interventions and coaching/mentoring students. There were thus effective interventions also within these subgroups of components, and all have at least one study with an average effect size larger than 0.10.

Personnel development, increased resources, and computer-assisted instruction have higher effect sizes, but only increased resources is statistically significant. The other two have wider confidence intervals, and there are interventions using these components with both relatively large positive and negative effect sizes. Content changes and coaching/mentoring are both statistically significant, and have effect sizes around 0.15 to 0.20. *Cooperative learning* and *small-group instruction* have larger effect sizes, both of which are significant. *Feedback and progress monitoring* and *tutoring* have the two largest effect sizes, both of which are significant. Only five studies include feedback and progress monitoring, and in all cases, this component is combined with at least one other component.

The I^2 statistics in Table 3 indicate that the components often account for more variation in effect sizes than the previously examined moderators, but there is still substantial heterogeneity. With the exception of summer programs, the Q test rejects the hypothesis of homogeneous effect sizes for all components that have been examined by more than four studies, and all prediction intervals except for increased resources and feedback and progress monitoring have a negative lower bound (test statistics not shown).

As almost all components have been examined by a relatively low number of studies, it was not possible to examine this heterogeneity further quantitatively for them. Thirty-six studies included a tutoring component and had at least one effect size measured at the end of intervention. To gauge the heterogeneity among studies with a tutoring-component, we conducted meta-regressions using only effect sizes from these studies as the dependent variable. We were unable to include more than three moderators (plus a constant) in the same regression before running into problems with degrees of freedom. All variables have missing observations, and we therefore used multiple imputation in all specifications.

In Table 4, we use the following moderators. Column (1) includes three variables relating to the type of participants. The share of girls is positive, and the share of minority and the share of low income students are negative, but all three are small and insignificant. Column (2) includes variables related to the delivery of the intervention, whether the intervention is delivered individually and whether it is delivered by professionals. Both variables are positively associated with effect sizes and nonsignificant, but the magnitudes indicate that they may still be important moderators.

Column (3) includes the duration of intervention and two other variables related to the dosage of an intervention that were not included earlier, namely,

TABLE 4

Results from meta-regressions on tutoring interventions examining differences in effect sizes attributable to moderators

Moderator	(1)	(2)	(3)	(4)
Participant characteristics				
Share of girls	0.00756			
	(0.00653)			
Share of minority	−0.000603			
	(0.003456)			
Share of low income	−0.000209			
	(0.00383)			
Intervention characteristics				
Individual recipient		0.125		0.111
		(0.0971)		(0.0904)
Delivered by professionals		0.0582		
		(0.0954)		
Duration in weeks			−0.00948**	−0.00906*
			(0.00365)	(0.00386)
Number of sessions			0.00246**	0.00190
			(0.00119)	(0.00128)
Hours per week			−0.0133	
			(0.0325)	
Constant	0.372**	0.283**	0.251**	0.210**
	(0.0513)	(0.0916)	(0.0495)	(0.0655)
Effect sizes (*n*)	202	202	202	202
Number of studies (*k*)	36	36	36	36
I^2	62.23	64.83	53.39	52.54

Note. Robust standard errors in parentheses. Missing values have been multiple imputed in all specifications.
$^*p < .05$. $^{**}p < .01$.

the frequency and intensity of delivery. Duration and frequency of delivery were significant; intensity was not significant. The coefficient on duration implies that 10 more intervention weeks is, on average, associated with a 0.09 smaller effect size. Frequency of delivery is measured in number of sessions, so a tutoring-intervention with 10 more sessions has on average a 0.02 larger effect size. Finally, Column (4) includes individual delivery, duration, and frequency jointly. The estimates become smaller for all variables, although the changes are not particularly large. Frequency ceases to be significant, whereas intervention duration remains significant. Further analysis indicates that six studies with a duration of 40 weeks—the mean and median duration for tutoring-interventions are around 18 to 19 weeks—have somewhat lower effect sizes. The weighted average for these six studies is still 0.21 and significant.

The I^2 decreases from Column (1) to (5), where it is 53%. But there is still some unexplained variation, and we cannot include all moderators in the same specification. In particular, study design (RCT) and tests conducted on subdomains would have been interesting to explore further. We have performed sensitivity analyses for these two variables, along with other robustness checks. These analyses are reported next.

Sensitivity Analysis

The sensitivity analysis focuses on the seven components for which we found significant weighted average effect sizes in the previous section: increased resources, content changes, coaching/mentoring personnel, cooperative learning, small-group instruction, feedback and progress monitoring, and tutoring.[11]

Study Design

The average effect sizes reported in Table 3 and Figure 5 include both RCTs and QES, but RCTs had lower effect sizes in our meta-regressions. The two studies of increased resources are large scale RCTs, however, and all studies of small-group instruction are QES. Three of the nine studies of content changes were QES and omitting these reduces the effect size from 0.19 to 0.15, which is still statistically significant. There are three studies of coaching/mentoring personnel, out of ten, that are QES. Removing these leaves the average effect size unchanged, but no longer significant. The effect size for cooperative learning is reduced from 0.22 to 0.10 when we remove the three of 10 QES studies. This weighted average effect size is still significant. Two of five studies of a feedback and progress monitoring component were QES. Omitting these studies lowers the effect size from 0.32 to 0.25, which is still significant. Five of 36 studies including a tutoring component were QES. When these studies are omitted the effect size increases slightly and the confidence interval moves further away from zero. Thus, for four of seven components, statistical significance does not rely on the inclusion of QES.

Outcome Measurement

Interventions using subtests of more general standardized tests had higher effect sizes in the meta-regressions. Our aim here is to explore a possible explanation of the differences in effect sizes between components. If some types of interventions systematically target general skills, which may be harder to affect, that is one such explanation. There is no general pattern though; the average effects sizes are larger on general tests for some components and lower for others. However, for the five components with the largest effect sizes, all average effect sizes are larger when subtests have been used. Content changes and coaching/mentoring personnel have relatively small and not statistically significant effect sizes using only general tests (0.08 and -0.02, respectively). Tutoring and feedback and progress monitoring are still significant using only general tests, and the lower bound on the 95% confidence interval for cooperative learning is −0.001.

Clustered Treatment Assignment

When studies used clustered assignment of treatment, we corrected effect sizes and standard errors where appropriate. The small-group instruction-category does not contain any interventions with clustered assignment of treatment. All other components are relatively robust to this adjustment, both in terms of average effect sizes and significance. The largest effect size change is for coaching/mentoring personnel (−0.017), and except for increased resources, all other components remained statistically significant.

Discussion

This systematic review has examined interventions that aim to improve the educational achievement for low SES students in elementary and middle school. The meta-analysis included 101 studies with a treatment-control design, 76% of which were RCTs, and where outcomes were measured by standardized tests in reading and mathematics. The analysis showed that there are interventions that improve the educational achievement of low SES students, and also that there is substantial variation in effect sizes. Moreover, observable study characteristics cannot fully explain this variation. In other words, there is still much to learn about how to design interventions for low SES students.

The examination of intervention components showed that tutoring, feedback and progress monitoring, and cooperative learning have comparatively large and robust average effect sizes. Tutoring in particular has a strong research base comprising many RCTs, and the effect was significant when included in a meta-regression. All other components, including feedback and progress monitoring and cooperative learning, have been examined in too few studies for meta-regressions to be conducted. Therefore, we do not know if and how these components are moderated by study characteristics.

Although the magnitudes of the average effect sizes for tutoring (0.36), feedback and progress monitoring (0.32), and cooperative learning (0.22) are not large enough to close the gap between high and low SES students, they represent a substantial reduction of that gap if targeted toward low SES students.[12] Using similar standardized tests as our review, Lipsey et al. (2012) reported reading (and mathematics) gaps in the United States between students with high and low SES of 0.74 (0.85) in Grade 4, and 0.66 (0.80) in Grade 8. Using another of Lipsey et al.'s metrics, effect sizes of 0.2 to 0.4 represent 13% to 26% of the average yearly improvement from Grades K to 1, and 83% to 167% of the yearly improvement from Grades 8 to 9. The duration of interventions in our review was on average well under a year, and less than half a year for tutoring, so the substantial improvements occurred in a relatively short period of time. Some of our findings are in line with recent reviews of interventions for other student populations. Slavin and Lake (2008), Slavin, Lake, and Groff (2009), and Slavin, Lake, Chambers, et al. (2009) also concluded that there is still a lot to be learned about the design of effective interventions. They also pointed to instructional-process programs as having the highest effect sizes for general student populations. Instructional-process programs include tutoring and cooperative learning as well as for example classroom management and motivation interventions, and teaching of metacognitive strategies. Tutoring and cooperative learning also figured prominently in Slavin et al. (2011), who reviewed programs for struggling readers.

Studies of feedback and progress monitoring in our material mainly give teachers more knowledge about student progression to allow teachers to adjust material and instruction appropriately. It should be emphasized that there are only five studies using this component, and in all cases this component is combined with another component. A similar type of intervention component has only been included in one other related review: feedback to teachers had significant effects on mathematics outcomes in Gersten et al. (2009). Feedback also seems to be an important part of effective instruction, in general (e.g., Hattie & Timperley, 2007). Recent experiments not included in our review, where teachers get help to individualize instruction by an algorithm-guided feedback program (e.g., Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Connor et al. 2013), or where districts and schools receive help with implementing data-driven development (Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013), also support the conclusion that feedback and progress monitoring can have positive effects for low SES students.[13]

Effect sizes of small-group instruction, coaching/mentoring of school personnel, content changes, and increased resources were also positive and significant. However, they were less robust in our sensitivity analysis. Small-group instruction, coaching/mentoring of school personnel, and increased resources were not robust to excluding QES studies, and all four were no longer significant when we restricted the sample to studies using tests of more than one subdomain of reading or mathematics. One reason may be that none of these components have been examined by a large number of studies. Thus, further research is needed. We have also found intervention components that on average have lower and nonsignificant effect sizes: Incentive programs, after-school programs, summer programs, coaching and mentoring of students, psychological/behavioral interventions, personnel development, and computer-assisted instruction programs were all positive but not significant. However, there are examples of effective interventions in many of these categories, and most of these components have been examined by only a few studies. Therefore, we want to emphasize that our results do not imply that incentive programs cannot be effective, or that there are no incentive programs with positive effects (there are such examples).

We have found few consistently significant moderators of effect sizes. Interventions measuring outcomes by testing subdomains of mathematics and reading have higher effect sizes compared with those using more general tests. A potential explanation is that more specific tests are closer to the content targeted by the interventions, but it should be mentioned that the pattern of higher effect sizes for tests of subdomains is not completely general; the average effects sizes are larger on general tests for some intervention components. Furthermore, RCTs have lower effect sizes than QES. Although this is an interesting result, it is difficult to examine the reasons behind it with our data. It is possible that effect sizes in QES are biased upward, for example, due to selection of more motivated students or teachers into interventions. RCTs may be downward biased due to more severe problems with blinding of treatment status, or because teachers and school personnel are less accepting of random than nonrandom assignment of students to interventions. The type of interventions studied in RCTs may also be different from those studied in QES.[14]

The share of minority students was significantly associated with lower effect sizes in our full specification, but the estimate is relatively small, highly correlated with the share of low income students (which is positively associated with effect sizes), and relies on imputed values. Furthermore, for tutoring studies we found a very small and nonsignificant negative association.

For the full set of effect sizes, there were no other significant variables related to type of study, measurement of effect sizes, participants as well as dosage and delivery of the interventions. For effect sizes from tutoring-interventions, the duration of the intervention was negatively and significantly associated with the effect size, a result that seemed to be driven by interventions with the longest duration (about 1 school year). The negative estimate may seem counterintuitive, but has been noted in other reviews as well (e.g., Wanzek et al., 2006; Wanzek et al., 2013). The small number of studies of tutoring interventions with long duration precludes a deeper exploration of the reasons for the negative association, but although the studies do not stand out regarding observed variables, such as participants and intensity, the interventions may have targeted a population that was different in some dimension we did not record.

Limitations

There are several limitations to our results. We focused on standardized test scores, as these tests tend to be less inherent to treatment (Slavin & Madden, 2011), and have been shown to more stable in comparisons over time (Scammaca et al., 2015). However, studies that were excluded because they did not use standardized tests may be of high quality and still provide useful information about how interventions affect students.

All tests that were included in the meta-analysis measured knowledge of reading or math, but these subjects have multiple dimensions. By grouping all tests into only two subject categories, we may have missed aspects of what the interventions were trying to achieve (although this was a necessary step in order to synthesize the results). The results in Scammaca et al. (2015) indicated that grouping tests is not likely to be a major problem for reading outcomes. They examined reading interventions directed to struggling readers and compared, among other things, whether they targeted reading comprehension, fluency, word study, vocabulary, or combinations of these components. For standardized tests, the only significant difference was a greater effect for word study interventions over fluency interventions. Potentially though, more knowledge could be gained by coding studies over both how the instructional methods and how the content of instruction is changed by interventions as well as their combinations. To achieve sufficient statistical power to detect such differences between intervention modes would require a substantial increase in the number of studies.

Several studies were not used in the meta-analysis because they lacked the information needed to calculate an effect size, and we were unable to retrieve the information from study authors. We had to limit the scope of the search by requiring that studies contained keywords matching the four dimensions demarking our topic: students, socioeconomic status, outcomes, and study design. A broader search would have included all studies that matched any of these dimensions, and it is possible that we missed relevant studies because of this. It is however unlikely that we missed essential parts of the literature.

Low SES students seem to have fewer resources in many domains. It is possible that interventions that combine components addressing different domains may be more effective than single component interventions. Combinations were relatively rare among the studies in this review (19 studies in total), however, and the number of repeated combinations were very few. We have therefore not been able to examine this issue. The dominance of U.S. studies may be a caveat for the generalizability of our results. However, few of the examined interventions should be dismissed out of hand as not possible to implement in other countries or contexts. Most seem compatible with any type of educational system or school.

Last, we deviated from the protocol in some ways. The protocol specified a broader scope. We undertook a re-scoping of the review to focus on elementary and middle school interventions only and standardized tests only. To operationalize the criteria that interventions should be aimed at low SES students, we considered a study for inclusion if at least 50% of the sample were low SES as reported in the study. The three moderators related to outcome assessment were added at the coding stage, as variation in measurement was identified during this process. One of the prior objectives for this review was to examine the costs of interventions systematically. Unfortunately, only a small number of studies reported intervention costs. We were similarly interested in examining longer run outcomes of interventions, but few studies reported such outcomes.

Implications for Practice and Research

Our results indicate that it is possible for schools and local stakeholders to substantially improve the educational achievement of low SES students. As such, they provide motivation for action. We hope that the review may provide inspiration for educational decision makers at all levels who are looking for ways to improve the educational achievement of low SES students. When it comes to the specific interventions a school or a teacher should choose, our results provide less guidance. Even in our large sample of studies, it was not possible to fully explain why some interventions worked better than others. We also believe that the impact of an intervention depends on the local context. Thus, the review can be a source of knowledge and inspiration to schools and local stakeholders, but it does not provide a complete blueprint for how to increase the performance of low SES students.

The results of this review also have several implications for education research. Almost all studies were from the United States. The evidence base for (or against) particular interventions was thin in most countries. At the same time, countries are emphasizing the need to improve educational achievement for disadvantaged students. Examining the effects of interventions targeting disadvantaged students and examining interventions outside the United States should be important tasks in the coming years. Both math and reading interventions were included in the meta-analysis, but the evidence base is much smaller for math. More interventions targeting math would therefore be a useful addition to the literature. The negative associations between effect sizes for RCTs and share of minority students will be interesting to study in future reviews and in primary research. The lack of estimates of intervention costs constitutes a serious limitation of the literature. Effect sizes cannot be the sole basis for choosing among interventions. Preferably, cost-effective interventions with at least moderate effect sizes should be chosen. Researchers should therefore strive to include at least an estimate of the costs of implementing an intervention. Similarly, to fully evaluate the cost-effectiveness of interventions, estimates of costs for longer-lasting effects are necessary.

Concluding Remarks

There are still many unknowns about providing best interventions to low SES students. Of particular concern to practitioners, policymakers, and researchers alike is the limited knowledge about the costs of interventions, and the scarcity of effect sizes measured farther away from the end of interventions. To design more effective interventions and implement existing ones better, intervention designers and school managers will benefit from more knowledge about moderators of effect sizes. However, the review has clearly shown that there are interventions capable of substantially improving educational achievement of students from families with low SES. Tutoring, cooperative learning, and feedback and progress monitoring seem to be especially promising components of such interventions. In sum, the results of this review provide motivation to increase efforts both to implement interventions for low SES students, and to design studies that may answer the pending questions of what types of interventions are most cost-effective.

Notes

Parts of the material used in this article were also used in *Skolerettede indsatser for elever med svag socioøkonomisk baggrund: En systematisk forskningskortlægning og syntese* (Dietrichson, Bøg, Filges, & Klint Jørgensen, 2015; SFI report 15:07; http://www.sfi.dk/sfi-rapporter-3278.aspx). The current version includes 101 studies, compared with 69 in the Danish report, and all analyses performed in the article are new. We are grateful to the Danish Ministry of Education for providing support for this research, and to Agi Csonka, Mette Deding, Ulla Højmark Jensen, Chantal Nielsen, Eva Jul Stenholt, Rikke Eline Wendt, and participants at the SFI Advisory Board Conference for helpful discussions and comments. Ole Gregersen, Rasmus Henriksen Klokker, Ulrik Højmark Pedersen, Trondur Møller Sandoy, Louise Kristiansen, Mia Thuelund Hansen, and Bjørn Viinholt Nielsen provided excellent research assistance.

[1] For a nuanced discussion of the interplay between nature and nurture in child development, see Rutter (2006).

[2] See Cook et al. (2014) for an example of a successful high school intervention building on the idea that tutoring and social–cognitive training components are complements, and Rossin-Slater and Wüst (2016) for an example where a health and a childcare intervention are substitutes.

[3] None of the reviews mentioned in the section Previous Reviews contain a systematic cost-effectiveness assessments.

[4] We have used the Stata command *robumeta*. The method requires an initial estimate, $\hat{\rho}$, of the correlation between treatments and tests within the same study. We have used $\hat{\rho} = 0.8$ (as, e.g., Hedges et al., 2010, and Wilson et al., 2011). The results are virtually unchanged if we instead use $\hat{\rho} = 0.7$ or 0.9. Another way of dealing with dependence would be to estimate hierarchical linear models. Scammaca, Roberts, and Stuebing (2014) compared the robust variance estimation method of Hedges et al. (2010) to a three-level meta-analysis model and found that they yielded similar results.

[5] See, for example, Rubin (1996) and Pigott (2009) on why leaving out studies/effect sizes with missing values normally yields biased estimates. We have used the Stata command *mi impute* with sequential imputation using chained equations to generate values for missing observations. The dependent and all independent variables without missing observations have been used in the estimation to impute values, except for the indicators for Glass's Δ and whether intervention provider receive training, which cause perfect predictions.

[6] We have also included one more study that only reported results from a composite mathematics and reading test in the regressions.

[7] We have included an indicator for tutoring interventions, but cannot include indicators for other components due to an insufficient number of studies.

[8] We have also coded whether students were the primary recipients of an intervention and whether the intervention was performed in school. There were very few interventions that do not have students as recipients, or are performed outside schools. Due to the lack of variation, we have omitted these two variables.

[9] This is the largest pairwise correlation in our sample. For other variables there should be little risk of multicollinearity, as the second largest pairwise correlation is 0.54 (between tutoring and testing on a subdomain).

[10] No component, except tutoring, has more than one study with a follow-up measure. There are six tutoring-studies that report follow-up measures. The weighted average effect size of these is 0.21, which is significant ($p < .05$).

[11] In addition to the analyses reported in this section, we evaluated publication bias for all studies and for tutoring studies separately, using funnel plots. There were indications of publication bias when we included all studies, but no clear indications for tutoring studies. All other components were evaluated in too few studies to be examined in this way.

[12] Our data do not tell us the effect sizes of such interventions for high-SES students, so if given to all students, the effect on the achievement gap is ambiguous.

[13] The samples in these studies included more than 50% low SES students. In Connor et al.'s (2007, 2013) studies, control group teachers received extensive professional development, making the study a comparison between two alternative interventions. The effect sizes in these studies were therefore not comparable to designs with a control group. In Slavin et al. (2013) treatment was assigned and implemented at the school district level, an assignment level not otherwise present in our material.

[14] Anglemyer, Horvath, and Bero (2014) examined differences between RCTs and observational studies for a wide range of health care outcomes in 14 systematic reviews and found no significant differences. We are not aware of a similar large scale study of educational outcomes, but Gersten et al. (2009) found smaller effect sizes of mathematics interventions interventions in RCTs compared with QES, whereas de Boer et al. (2014) found the opposite examining learning strategy instruction interventions. The differences were not significant in either study.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Education Evaluation and Policy Analysis*, *23*, 171–191. doi:10.3102/01623737023002171

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*, 1–18. doi:10.1037/a0021017

Andreassen, C., & Fletcher, P. (2007). *Early Childhood Longitudinal Study, Birth Cohort (ECLS–B) psychometric report for the 2-year data collection* (Institute of Education Sciences Report; NCES 2007-084). Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007084

Anglemyer, A., Horvath, H. T., & Bero, L. (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews*, (4), MR000034. doi:10.1002/14651858.MR000034.pub2

Björklund, A., & Salvanes, K. (2011). Education and family background. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 201–247). Amsterdam, Netherlands: North-Holland.

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

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*, 125–230. doi:10.3102/00346543073002125

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371–399. doi:10.1146/annurev.psych.53.100901.135233

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*, 2593–2632. doi:10.1257/aer.104.9.2593

Cheung, A. C. K., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review*, *7*, 198–215. doi:10.1016/j.edurev.2012.05.002

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, *24*, 1408–1419. doi:10.1177/0956797612472204

Connor, C. M., Morrison, F. J., Fishman, B., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, *315*, 464–465. doi:10.1126/science.1134513

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. J., Guryan, J., Ludwig, J., . . . Steinberg, L. (2014). *The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in Chicago* (NBER Working Paper No. 19862). Retrieved from http://www.nber.org/papers/w19862

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood and human capital development. *Journal of Economic Literature*, *47*, 87–122. doi:10.1257/jel.47.1.87

de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research*, *84*, 509–545. doi:10.3102/0034654314540006

Dexter, D. D., & Hughes, C. A. (2011). Graphic organizers and students with learning disabilities. *Learning Disability Quarterly*, *34*, 51–72. doi:10.1177/073194871103400104

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Klingler Tackett, K., & Wick Schnakenberg, J. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research*, *79*, 262–300. doi:10.3102/0034654308325998

Epple, D., Romano, R., & Zimmer, R. (2015). *Charter schools: A survey of research on their characteristics and effectiveness* (NBER Working Paper No. 21256). Retrieved from http://www.nber.org/papers/w21256

Esping-Andersen, G., Garfinkel, I., Han, W.-J., Magnuson, K., Wagner, S., & Waldfogel, J. (2012). Child care and school performance in Denmark and the United States. *Children and Youth Services Review*, *34*, 576–589. doi:10.1016/j.childyouth.2011.10.010

Flynn, L. J., Zheng, X., & Swanson, H. L. (2012). Instructing struggling older readers: A selective meta-analysis of intervention research. *Learning Disabilities Research & Practice*, *27*, 21–32. doi:10.1111/j.1540-5826.2011.00347.x

Fryer, R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *Quarterly Journal of Economics*, *129*, 1355–1407. doi:10.1093/qje/qju011

Fryer, R. G., & Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review*, *103*, 981–1005. doi:10.1257/aer.103.2.981

Gershenson, S. (2013). Do summer time-use gaps vary by socioeconomic status? *American Educational Research Journal*, *50*, 1219–1248. doi:10.3102/0002831213502516

Gersten, R., Chard, D. J., Jayanti, M., Baker, S. K., Morphy, P., & Flojo, P. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*, 1202–1242. doi:10.3102/0034654309334431

Glazerman, S., Protik, A., Teh, B. R., Bruch, J., & Max, J. (2013). *Transfer incentives for high-performing teachers. Final results from a multisite randomized experiment* (Institute of Education Sciences Report; NCEE 2014–4003). Retrieved from http://ies.ed.gov/ncee/pubs/20144003/pdf/20144003.pdf

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*, 645–662. doi:10.1016/j.appdev.2003.09.002

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*, 183–208. doi:10.1007/s11881-010-0041-x

Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Sciences*, *13*, 65–73. doi:10.1016/j.tics.2008.11.003

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*(1), 4–9. Retrieved from https://www.aft.org/sites/default/files/periodicals/TheEarlyCatastrophe.pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. doi:10.3102/003465430298487

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902. doi:10.1126/science.1128898

Hedges, L. V. (2006). Meta-analysis. In C. R. Rao (Ed.), *The handbook of statistics* (Vol. 26, pp. 919–953). Amsterdam, Netherlands: Elsevier.

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

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Education Evaluation and Policy Analysis*, *29*, 60–87. doi:10.3102/0162373707299706

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

Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2015). *Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago* (NBER Working Paper No. 21178). Retrieved from http://www.nber.org/papers/w21178

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). Retrieved from http://www.cochrane-handbook.org

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560. doi:10.1136/bmj.327.7414.557

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*, 386–431. doi:10.3102/0034654313483906

Kyndt, E., Raes, E., Lismont, B., Timmers, F., Cascallar, E., & Dochy, F. (2013). A meta-analysis of the effects of face-to-face cooperative learning. Do recent studies falsify or verify earlier findings? *Educational Research Review*, *10*, 133–149. doi:10.1016/j.edurev.2013.02.002

Lipsey, M. W., Puzio, K., Yun, C., Herbert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (Institute of Education Sciences Report; NCSER 2013-3000). Retrieved from https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage.

Magnuson, K., & Shager, H. (2010). Early education: Progress and promises for children from low-income families. *Children and Youth Services Review*, *32*, 1186–1198. doi:10.1016/j.childyouth.2010.03.006

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence—New findings and theoretical developments. *American Psychologist*, *67*, 130–159. doi:10.1037/a0026699

Organisation for Economic Co-operation and Development. (2010). *PISA 2009 results: Overcoming social background: Equity in learning opportunities and outcomes* (Vol. II). Paris, France: Author. doi:10.1787/9789264091504-en

Organisation for Economic Co-operation and Development. (2013). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed* (Vol. II). Paris, France: Author. doi:10.1787/9789264201132-en

Pigott, T. D. (2009). Handling missing data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 399–416). New York, NY: Russell Sage Foundation.

Reljic, G., Ferring, D., & Martin, R. (2015). A meta-analysis on the effectiveness of bilingual programs in Europe. *Review of Educational Research*, *85*, 92–128. doi:10.3102/0034654314548514

Reynolds, A. J., Magnuson, K. A., & Ou, S-R. (2010). Preschool-to-third grade programs and practices: A review of research. *Child and Youth Services Review*, *32*, 1121–1131. doi:10.1016/j.childyouth.2009.10.017

Rhemtulla, M., & Tucker-Drob, E. M. (2012). Gene-by-socioeconomic status interaction on school readiness. *Behavioral Genetics*, *42*, 549–558. doi:10.1007/s10519-012-9527-0

Ritter, G., Albin, G., Barnett, J., Blankenship, V., & Denny, G. (2006). The effectiveness of volunteer tutoring programs: A systematic review. *Campbell Systematic Reviews*, *2*(7). doi:10.4073/csr.2006.7

Robinson, D. R., Ward Schofield, J., & Steers-Wentzell, K. L. (2005). Peer- and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review*, *17*, 327–362. doi:10.1007/s10648-005-8137-2

Rossin-Slater, M., & Wüst, M. (2016). *Are different early investments complements or substitutes? Long-run and intergenerational evidence from Denmark* (Unpublished manuscript). Department of Economics, University of California at Santa Barbara, CA. Retrieved from http://econ.ucsb.edu/~mrossin/RossinSlater_Wust_interactions_mar2016.pdf

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489. doi:10.1080/01621459.1996.10476908

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*, 79. doi:10.1186/1471-2288-8-79

Rutter, M. (2006). *Genes and behavior: Nature-nurture interplay explained*. Malden, MA: Blackwell.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*, 448–467. doi:10.1037/1082-989X.8.4.448

Scammaca, N. K., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*, 328–364. doi:10.3102/0034654313500826

Scammaca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*, 369–390. doi:10.1177/0022219413504995

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*, 417–453. doi:10.3102/00346543075003417

Slates, S. L., Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2012). Counteracting summer slide: Social capital resources within socio-economically disadvantaged families. *Journal of Education for Students Placed at Risk*, *17*, 165–185. doi:10.1080/10824669.2012.688171

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*, *50*, 371–396. doi:10.3102/0002831212466909

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*, 427–515. doi:10.3102/0034654308317473

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*, 1391–1466. doi:10.3102/0034654309341374

Slavin, R. E., Lake, C., Davis, S., & Madden, N. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*, 1–26. doi:10.1016/j.edurev.2010.07.002

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*, 839–911. doi:10.3102/0034654308330968

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

Timperley, H. S., & Phillips, G. (2003). Changing and sustaining teachers' expectations through professional development in literacy. *Teaching and Teacher Education*, *19*, 627–641. doi:10.1016/S0742-051X(03)00058-1

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

Tucker-Drob, E. M., Briley, D. A., & Harden, K. P. (2013). Genetic and environmental influences on cognition across development and context. *Current Directions in Psychological Science*, *22*, 349–355. doi:10.1177/0963721413485087

Tucker-Drob, E. M., Rhemtulla, M., Harden, K. P., Turkheimer, E., & Fask, D. (2011). Emergence of a gene x socioeconomic status interaction on infant mental ability between 10 months and 2 years. *Psychological Science*, *22*, 125–133. doi:10.1177/0956797610392926

UNESCO. (1994). *The Salamanca statement and framework for action on special needs education*. Retrieved from https://www.european-agency.org/sites/default/files/salamanca-statement-and-framework.pdf

Wanzek, J., Vaughn, S., Scammaca, N. K., Metz, K., Murray, C. S, Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*, *83*, 163–195. doi:10.3102/0034654313477212

Wanzek, J., Vaughn, S., Wexler, J., Swanson, E. A., Edmonds, M., & Kim, A-H. (2006). A synthesis of spelling and reading interventions and their effects on the spelling outcomes of students with LD. *Journal of Learning Disabilities*, *39*, 528–543. doi: 10.1177/00222194060390060501

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*, 461–481. doi:10.1037/0033-2909.91.3.461

Wilson, S., Tanner-Smith, E. E., Lipsey, M. W., Steinka-Fry, K., & Morrison, J. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout among school-aged children and youth. *Campbell Systematic Reviews*, *7*(8). doi:10.4073/csr.2011.8

Zief, S. G., Lauver, S., & Maynard, R. A. (2006). Impacts of after-school programs on student outcomes: A systematic review. *Campbell Systematic Reviews*, *2*(3). doi:10.4073/csr.2006.3

Authors

JENS DIETRICHSON, PhD, is a researcher at the Schooling and Education & SFI Campbell Unit at SFI – The Danish National Centre for Social Research, Herluf Trollesgade 11, DK 1052 Copenhagen K, Denmark; email: *jsd@sfi.dk*.

MARTIN BØG, PhD, is a researcher at the Schooling and Education & SFI Campbell Unit at SFI – The Danish National Centre for Social Research, Herluf Trollesgade 11, DK 1052 Copenhagen K, Denmark; email: *martin.bog@gmail.com*.

TRINE FILGES, PhD, is a senior researcher at the SFI Campbell Unit at SFI – The Danish National Centre for Social Research, Herluf Trollesgade 11, DK 1052 Copenhagen K, Denmark; email: *tif@sfi.dk*.

ANNE-MARIE KLINT JØRGENSEN is an information specialist at the Schooling and Education & SFI Campbell Unit at SFI – The Danish National Centre for Social Research, Herluf Trollesgade 11, DK 1052 Copenhagen K, Denmark; email: *amk@sfi.dk*.