Journal of Research on Educational Effectiveness

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uree20

Cooperative Learning and Literacy: A Meta-Analytic Review

Kelly Puzio [a] & Glenn T. Colby [b]

[a] Washington State University , Pullman , Washington , USA

[b] University of Colorado Boulder , Boulder , Colorado , USA

Published online: 04 Oct 2013.

To cite this article: Kelly Puzio & Glenn T. Colby (2013) Cooperative Learning and Literacy: A Meta-Analytic Review, Journal of Research on Educational Effectiveness, 6:4, 339-360, DOI: 10.1080/19345747.2013.775683

To link to this article: http://dx.doi.org/10.1080/19345747.2013.775683

Journal of Research on Educational Effectiveness, 6: 339–360, 2013

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2013.775683

Routledge
Taylor & Francis Group

Cooperative Learning and Literacy: A Meta-Analytic Review

Kelly Puzio
Washington State University, Pullman, Washington, USA

Glenn T. Colby
University of Colorado Boulder, Boulder, Colorado, USA

Abstract: We conducted a meta-analysis on the effectiveness of cooperative and collaborative learning to support enhanced literacy outcomes. Interventions considered were provided in regular education settings (i.e., not pull-out instruction) with students from Grades 2 through 12. Reviewing more than 30 years of literacy research, we located 18 intervention studies with 29 study cohorts. Included studies primarily used standardized assessments to report on students' reading, vocabulary, or comprehension achievement, which we analyzed separately. Overall, students had significantly higher literacy achievement scores when instructional interventions utilized cooperative and collaborative activity structures. The overall weighted mean effect sizes ranged from 0.16 to 0.22 ($p < .01$) with more than 94% of the point estimates being positive. Because cooperative or collaborative learning was always one of multiple intervention components, it was impossible to estimate the unique, added effects of cooperative/collaborative learning. Although the small number of eligible studies precludes any claims about the effectiveness of specific forms of grouping and the circumstances under which programs have more impact, our findings suggest that cooperative and collaborative grouping was a core component of effective literacy interventions, particularly at the elementary level.

Keywords: Literacy instruction, meta-analysis, cooperative learning

BACKGROUND

Most children can—and do—learn to read. A wide variety of data, however, indicate that too many of our children are reading poorly. At Grades 4 and 8, only about one third of our nation's students are identified as proficient readers (National Center for Education Statistics, 2009). In a recent administration of the National Assessment of Educational Progress (National Center for Education Statistics, 2009), most students located and recalled text information, but they commonly failed to make sound inferences and evaluations of informational and narrative text. Recent international literacy assessments (e.g., OECD, 2010) have led to similar findings, and researchers have concluded that U.S. students, when compared to other developed nations, read at "average" levels (Fleischman, Hopstock, Pelczar, & Shelley, 2010). In light of these findings, we—as a country—need to carefully examine the ways in which educators organize reading instruction.

Teachers, educational specialists, schools, and district staff support children's reading by designing learning environments where students are involved in complex meaning

Address correspondence to Kelly Puzio, Washington State University, Teaching & Learning, 329 Cleveland Hall, Pullman, WA 99164. E-mail: kelly.puzio@wsu.edu

making. To this end, thoughtfully designed learning environments (Lehrer & Schauble, 2000) employ a wide variety of tasks (e.g., reading or writing prompts), tools (e.g., text audio recordings, text to speech software), texts (e.g., poems, graphic novels), modes and means of argumentation (e.g., question–answer, cite the text), and activity structures (e.g., whole-class instruction, individualized instruction). The practice of using small-group activity structures has a long history in the field of education (e.g., Evans, 1942; Shields, 1927). During this history, researchers have used a wide variety of grouping strategies (e.g., "ability" [1] groups, cooperative groups). As an important note, our focus in this article is within-class grouping, not between-class grouping (Gamoran, 1992), which is also known as tracking or streaming.

One particular form of within-class grouping that is widely known, heavily debated, and commonly researched is "ability" grouping. Researchers have argued that "ability" grouping widens the achievement gap (e.g., Hiebert, 1983; Oakes, 1985), provides inferior instruction for "low-ability" groups (e.g., Allington, 1983; Wheelock, 1994), and stigmatizes lower level groups (Borko & Eisenhart, 1989; Peterson, 1989). Over time, students in low-ability groups may lose self-esteem and motivation for learning (Reutzel & Cooter, 1991). In particular, ability groups are said to be especially harmful for students who have historically been marginalized because ability grouping perpetuates larger social inequities (Braddock & Slavin, 1992). In addition, researchers have argued within-class (Rosenbaum, 1976) "ability" grouping may serve to increase divisions along ethnic, racial, and class lines. The negative impact of grouping and admonishments not to use this practice has also received considerable coverage in current reading instruction textbooks and other venues (e.g., Baker, Dreher, & Guthrie, 2000).

In the wake of these negative claims, the practice of within-class grouping *in general* appears to be declining. In the 1960s through the 1980s, teachers reported using small-group activity structures in approximately 80% of U.S. elementary schools (Austin & Morrison, 1963; Weinstein, 1976). According to more recent surveys, approximately 60% of teachers report using some form of small-group activity (Baumann, Hoffman, Duffy-Hester, & Moon Ro, 2000; Chorzempa & Graham, 2006). A recent study systematically documented the literacy grouping practices of 184 fourth- and fifth-grade teachers using instructional logs and classroom observations and found that small-group activity structures were employed in only 30 to 40% of instructional time (Cordray, Pion, Brandt, & Molefe, 2011). Thus, a wide body of converging evidence suggests that the use of small-group activity structures is on the decline, particularly in elementary settings. Based on our review of the literature, there is not enough evidence to make claims about the amount of within-class grouping occurring in secondary settings.

Although the use of small-group activity structures appears to be declining in elementary school settings, educational researchers continue to argue that cooperative and collaborative activity structures *in particular* are effective at supporting learning. For example, quantitative research shows that cooperative (e.g., Slavin, Lake, Chambers, Cheung, & Davis, 2009) and collaborative (e.g., Hitchcock, Dimino, Kurki, Wilkins, & Gersten, 2011) learning support reading, comprehension, and vocabulary development. Although

[1] We use quotation marks around the term "ability" to indicate that we question the way that this body of literature uses this term. In this corpus—and historically—the term problematically suggests that intelligence is a fixed trait. Similarly, we also place the terms "heterogeneous" and "homogeneous" in quotation marks to signal that these terms oversimplify student diversity. Although these terms provide some information about how groups were formed, we believe that every group is heterogeneous and diverse when student culture, history, language, and literacy are considered deeply.

"cooperative" and "collaborative" learning are generic terms that describe a host of programs and interventions, both activity structures organize student learning by inviting small groups of students to participate in goal-directed activities and to interact verbally and socially. When compared with collaborative learning, cooperative learning is typically viewed as more structured, more prescriptive (e.g., can include teacher scripts), and more directive about how students work together (e.g., participation roles; see Oxford, 1997, for a detailed comparison). It is important to note that cooperative and collaborative learning are distinct from "ability" grouping insofar as small groups are designed to be academically "mixed" or heterogeneous.

PREVIOUS REVIEWS

Although several reviews have been conducted on classroom grouping, their analytic focus has varied widely. Early reviews that examined grouping focused most of their attention on the contentious debate on between-class grouping (Kulik, 1992; Kulik & Kulik, 1987; Slavin, 1987, 1990). Some previous reviews have focused exclusively on within-class grouping, but they have either collected and analyzed empirical studies with any outcome, such as grammar, mathematics, or science (Johnson & Johnson, 1986; Lou et al., 1996), or they have reviewed pull-out reading programs for special education students (Elbaum, Vaughn, Hughes, & Moody, 1999). Two recent reviews have collected and analyzed any reading program evaluation at the elementary (Slavin et al., 2009) and middle/high school (Slavin, Cheung, Groff, & Lake, 2008) levels. Although these reviews did not use advanced meta-analytic techniques, both reviews argued that the most successful reading programs have cooperative learning at their core.

Many reviews have analyzed the practice of within-class grouping, but none have focused on cooperative or collaborative learning for reading. Johnson and Johnson (1986) conducted a systematic meta-analysis of 578 studies on cooperative and competitive group learning, yet only 21 studies reported on any type of literacy outcome and most assessed grammar or spelling. Kulik and Kulik (1987) considered 19 within-class grouping studies, but only six studies reported literacy outcomes. Slavin's (1987) review considered eight within-class grouping studies, but only one of these studies reported a reading outcome. Citing mean effect sizes of +0.65, +0.27, and +0.41 for low, average, and high achievers, Slavin supported the practice of within-class "ability" grouping in upper elementary mathematics. Slavin also stated that there was not enough research on within-class "ability" grouping in reading to permit any conclusions. Analyzing 145 effect sizes for any academic outcome, Lou et al. (1996) reported an overall mean weighted effect size of +0.17 in favor of small-group instruction. Like previous reviews, however, their overall sample of effect sizes was composed mainly of mathematics outcomes. The last review on within-class grouping focused on reading for students with disabilities; Elbaum et al. (1999) reported a mean weighted effect size of +0.58 for within-class grouping studies, and the majority of these studies were cross-age tutoring dyads in pull-out special education programs.

Considering previous reviews on within-class and between-class grouping, only one took the additional step of analyzing effect size moderators. Lou, Abrami, and Spence (2000) used hierarchical linear modeling to explore effect size moderators and found outcome type (i.e., researcher or standardized test), teacher professional development, grouping specificity (groups formation strategy), type of small-group instruction (homogeneous or heterogeneous), grade level, and relative ability level (e.g., high achieving, low achieving)

to be statistically reliable predictors of the mean weighted effect size. As stated previously, however, the bulk of their effect sizes (> 60%) had mathematics outcomes.

REVIEW OBJECTIVES

The main objective of this review is to gather, summarize, and synthesize the empirical findings on the effects of cooperative and collaborative learning on literacy achievement. In particular, we focused on instructional programs implemented in the regular classroom setting—not learning specialists in pull-out settings where students leave the classroom. The goal of this analysis is to help policymakers, educators, parents, and other stakeholders understand the impact of designing literacy environments that use cooperative and collaborative learning principles. If the data support further analysis, we also want to know if any moderators (e.g., grade, outcome) influence the effectiveness of cooperative and collaborative learning. Our research questions are as follows:

1. To what extent do cooperative and collaborative literacy programs implemented in the general education classroom impact student literacy achievement?
2. Do any moderators (e.g., grade, outcome, teacher professional development) help explain this effect?

METHODS

Inclusion/Exclusion Criteria

All studies included in this review reported on the effects of cooperative or collaborative grouping on literacy (e.g., global reading, comprehension, vocabulary) achievement. Cooperative and collaborative programs were often combined with other educational programs, such as cognitive strategy instruction or a specific curriculum. This review focused on interventions delivered to school-aged children in Grades 2 through 12 (or equivalent grades in international settings) in regular classroom settings during school hours, so studies of college or adult students were not included. Studies of special education classrooms and alternative school settings were eligible for inclusion, but studies where the interventionist pulled students out of their regular classrooms for special instruction were not eligible. After-school programs were also not eligible. In the event of mixed-aged classrooms, partnering students, we included studies where students' age was within 1 year; therefore, studies that utilized cross-age peer interaction (e.g., Top & Osguthorpe, 1987) were not included. In keeping with previous reviews (e.g., Lou et al., 1996), interventions that specifically trained students in one-to-one peer tutoring (e.g., Jenkins et al., 1994) were not eligible. Between-classroom grouping studies—ones that placed children in different rooms for a semester or a year—were also not eligible. Studies conducted in any country were eligible, but only English-language studies were considered due to the limited resources for this review.

Only studies designed to make causal inferences were included. This included experimental studies and quasi-experimental studies that employed a pretest, posttest, and a counterfactual. Studies without control or comparison groups were not eligible. The majority of qualifying studies utilized a quasi-experimental design with post hoc matching,

but showed evidence of pretest equivalence on reading outcomes. Two fully randomized studies (Chamberlain, Daniels, Madden, & Slavin, 2009; Slavin, Chamberlain, Daniels, & Madden, 2009) met all inclusion criteria, and both included pretests. Finally, although the impact of grouping has been studied for many years (e.g., Evans, 1942; Shields, 1927), early studies reported very large effect sizes (greater than +1.0). To better approximate the impact that could be expected from cooperative and collaborative grouping intervention in today's educational context, we included only studies published in or after 1980.

Literacy Assessments

This review focused on reading, vocabulary, or comprehension achievement outcomes, so studies that reported only spelling or grammar outcomes were not included. Included studies most often utilized standardized assessments, but some researchers developed their own local assessments of reading, vocabulary, or comprehension. In addition, students were required to read the text or passages, so studies (e.g., Englert & Mariage, 1991) where an adult read the text orally to the student were not included. In this sample, the most common assessment was the California Achievement Test ($n = 11$).

Search Strategy

For this review, we identified potentially eligible studies using several methods. First, we conducted two separate searches in Cambridge Scientific Abstracts using the following databases: ERIC (Education Resources Information Center), International Bibliography of the Social Sciences, PsycARTICLES, and PsycINFO. The first Cambridge Scientific Abstracts search (see Table 1) attempted to locate studies on cooperative and collaborative learning with reading outcomes, and the second search attempted to locate studies on two particular literacy practices known to employ grouping—guided reading and jigsaw reading. Second, we inspected the bibliographies of relevant reviews, and, when a study was deemed eligible, we inspected its bibliography for additional candidate studies. Third, we searched ProQuest Digital Dissertations for unpublished dissertations. Although we

Table 1. Search terms and results

Search Source	Search Terms	Results
CSA	Descriptor = "reading comprehension" AND Abstract = group* or coop* or collab*	3,058
CSA	Abstract = "guided reading" OR jigsaw AND read*	459
Related Reviews	Elbaum (1999); Johnson and Johnson (1989); Kulik (1992); Kulik & Kulik (1987); Lou et al. (1996); Slavin (1987, 1990); Slavin et al. (2008)	1,476
Proquest Digital Dissertations	Index Term = "reading comprehension" AND Abstract = coop* OR collab* OR group*	199
Total		5,192

Note. CSA = Cambridge Scientific Abstracts.

found multiple dissertations that investigated cooperative or collaborative learning (e.g., Rapp, 1991), these studies typically provided pull-out instruction and thus did not meet our inclusion criteria. Table 1 details the operational definition of the search, the search results, and the overall search sequence and strategy.

For each potentially eligible study, the abstract was first read to determine whether the study met the inclusion/exclusion criteria previously stated. During this initial screening round, we eliminated studies based on the following guidelines:

- The study reported no literacy outcome
- The study did not employ at least a quasi-experimental pre/post control group design (e.g., case study, narrative, or qualitative study)
- The study used peer tutoring
- The students fell outside the parameters of Grades 2 to 12
- The study was not delivered in regular classroom settings (e.g., pull-out program)

For a study to be eliminated, the abstract had to include information clearly indicating that the study met one of the exclusion criteria just described. If not enough information was provided, the study was passed along to the next round for further analysis. The majority of articles during this round were excluded for the following reasons: The studies did not employ at least a quasi-experimental design, did not have an ungrouped control, or the studies' participating students fell outside the parameters of Grades 2 to 12. Only 235 studies passed this initial screening.

Both authors conducted full-text reviews of the remaining 235 articles. The initial evaluation focused on the introduction and Methods section. The purpose of this step was to achieve a fuller understanding of each study's scope and design. First, the reviewer read the results section to ensure that empirical results were reported—not simply program description—and that a reading outcome was measured and reported. Second, the Methods section was reviewed to ensure that the study employed a quasi-experimental or experimental design. This stage also resulted in a substantial narrowing of the study pool. Only 80 articles passed this stage and were subsequently fully read and coded. The full reading stage substantially narrowed the pool of candidate studies. Throughout the entire process, articles were excluded for a variety of reasons: no literacy outcome ($n = 104$), a design that would not produce an effect size ($n = 57$), peer tutoring program ($n = 20$), pull-out program ($n = 16$), participants not in Grade 2 to 12 ($n = 12$), and other reasons ($n = 7$). In the end, 18 unique studies were included in this review.

After these 18 studies were identified as having fully met our inclusion criteria, we analyzed the publication sources of these 18 studies. The top three journals that published these studies were *The Elementary School Journal, The Journal of Educational Psychology,* and *Reading Research Quarterly*. In an effort to locate studies missed by our search strategy, we retrieved and reviewed abstracts of every article published after 1980 in these three journals. No additional studies were found using this approach.

Data Management, Coding, and Reliability

Of the 235 full-text articles reviewed, 157 articles were available from online sources; 26 articles were available in the library on microfiche, microfilm, or in bound periodicals; and 52 articles (and dissertations) were acquired through interlibrary loan. Each reviewer separately coded the studies included in this review using the following categories:

- Study citation, author affiliation, type of publication, country
- Student characteristics (age, grade, race, sex)
- Sampling and assignment procedure
- Pretest differences and pretest equivalence
- Intervention and control sample size (start and finish)
- Intervention characteristics (e.g., professional development, intensity, duration, group size, ability level)
- Moderators suggested by previous meta-analyses (outcome type, teacher professional development, group specification, type of small-group instruction, relative ability level)
- Results and effect sizes

A preestablished coding form was used. Including effect size statistics, a total of 50 variables were coded for each study. As mentioned in the previous list, we specifically coded empirical studies for the six statistically reliable moderators reported by Lou et al. (2000) previously mentioned. Study coding and data management was initially done using Microsoft Excel; after all coding was complete, the data were transferred to SPSS and Comprehensive Meta-Analysis (CMA) for further analyses.

All reports were double-coded for effect size statistics and study characteristics. Discrepancies were discussed one at a time until a consensus was reached. Overall, the coding reliability was satisfactory. The mean kappa coefficient on categorical moderators was 0.86 with a minimum value of 0.41 and a maximum value of 1.0. In cases of disagreement, the coders discussed each item to resolution. Both authors coded the effect size statistics for every study. Coder reliability was determined by calculating a Pearson's correlation (r) for each statistic required to calculate an effect size (e.g., sample size, standard deviation, intervention mean). Pearson's correlation for the effect size statistic was 0.89. Reliability on other effect size data ranged from 0.89 (minimum) to 0.99 (maximum).

Statistical Procedures

We used Hedges's unbiased estimate (Hedges's g) of the standardized mean difference effect size statistic (the difference between the treatment and control group means on an outcome variable divided by the pooled standard deviations for the posttest measure). When available, we computed effect sizes using covariate adjusted posttest means. When these were unavailable, we subtracted pretest means from the posttest means to adjust for baseline differences. We used raw (i.e., unadjusted) group means only as a last resort. In all cases, we used the pooled posttest (unadjusted) standard deviation as the effect size denominator.

To create sets of independent effect size estimates for this analysis, we used only one effect size from each "study cohort" in any analysis. The study cohort term recognizes that many studies report effect size data separately for multiple samples, such as Grade 3, 4, and 5. In instances where more than one reading outcome was reported (e.g., more than one standardized reading assessment), the assessment most frequently used in the study corpus was used in the final analysis.

Because almost every study implemented an intervention to a naturally occurring classroom of students, observed sample sizes were adjusted to account for clustering (Hedges, 2004a, 2004b, 2007). When assignment to treatment or control is made at the aggregate level, the observed sample size is not appropriate for calculating the variance because students in naturally occurring classrooms cannot be viewed as statistically independent. Standardized mean difference effect sizes, observed samples sizes, and standard errors from

such studies were adjusted using a Microsoft Excel cluster adjustment calculator (McHugh, 2004), which uses algorithms based on suggestion of Hedges (2007). This adjustment requires an intraclass correlation coefficient (ICC). Because studies did not report ICCs, we used a conservative estimate (0.2) as an ICC approximation based on the findings of Hedges and Hedberg (2007), who concluded that the mean (across grade level) ICC for reading outcomes was 0.224. Although these cluster adjustments have minimal effect on the effect size statistic, they substantially reduce the effective sample size, which influences the effect size standard errors and the inverse variance weighting.

After we made the appropriate adjustments for clustering, we used Hedges and Olkin's (1985) small sample size adjustment to calculate an unbiased estimate of the population effect size:

$$\text{Hedges}' \; g = \left[1 - \frac{3}{4N - 9}\right] ES_{sm}$$

N is the total cluster-adjusted sample size ($n_{G1} + n_{G2}$) and ES_{sm} is the standardized mean difference effect size just discussed. When calculating overall effect sizes, we weighted each effect size by its inverse variance in all computations so that its contribution was proportionate to its reliability (Hedges & Olkin, 1985). We used a random effects statistical model to analyze effect sizes throughout this meta-analysis because we expected significant variability across effect sizes and because we did not want to restrict generalization of the findings only to the specific studies located for the analysis (Hedges & Vevea, 1998). The standard error for ES_i (that is, ES_i) and the sampling variance (SE_i^2) and, therefore, the inverse variance weight $\left(1/SE_i^2\right)$ are defined differently for the fixed effects and random effect analysis models. For the random effects analysis model, we used the following formulas:

$$SE_i^2 = v_i = v_i + v_\theta$$

$$v_i = \frac{n_t + n_c}{n_t n_c} + \frac{ES_i^2}{2(n_t + n_c)}.$$

The variable n_t is the observed posttest sample size in the treatment group, and n_c is the observed posttest sample size in the control group. V_θ was computed as appropriate in each analysis by the CMA software program or SPSS. We examined the cluster-adjusted sample size and Hedges's g effect size data for outliers using Tukey's (1977) definition of any point 3×IQR or more above the third quartile or 3×IQR or more below the first quartile; we found no outliers.

After these calculations, the Hedges's g statistics and their associated standard errors were input into CMA for each study. At this point, we tested the full collection of effect sizes for heterogeneity using the Q-statistic test (Hedges & Olkin, 1985). This test assumes the null hypothesis of homogeneity, where the differences among effect sizes are assumed to be due to subject-level sampling error. Although a multiple regression of effect size moderators was planned, there was not enough heterogeneity among the effect size statistics to warrant this analysis. This analysis would have capitalized on chance rather than true variation among the effect sizes.

RESULTS

Description of Included Studies

This review included 18 unique studies (14 journal articles and four technical reports). The majority of studies were conducted in urban or suburban elementary schools in the United States. Student participants represented a wide variety of histories, backgrounds, and cultures. Most studies lasted between 20 and 40 weeks and utilized a quasi-experimental design with a nonrandomly assigned matched comparison as the counterfactual condition. Every intervention study was delivered to full, intact classrooms, where classes, grades, or schools were assigned to receive the intervention program. In all but one case (Klinger, Vaughn, & Schumm, 1998), the classroom teacher was the primary instructor. Table 2 summarizes other characteristics of the included studies. A wide variety of cooperative and collaborative literacy programs were included in this review. The two most common programs are described here.

Five empirical studies researched the effectiveness of Cooperative Integrated Reading and Composition (CIRC). CIRC is a reading and writing program for students in Grades 2

Table 2. Characteristics of included studies

Characteristic	N	Characteristic	N
Publication Year		**Race/Ethnicity** *(Predominant)*	
1980s	5	White	4
1990s	8	Black	1
2000s	5	Hispanic	3
		Mixed	5
Publication		Lebanese	1
Journal article	15	Cannot Tell	4
Conference paper or technical paper	3		
		Location of study	
Study Design		USA	17
Quasi-experiment	10	Outside USA	1
Matched comparison quasi-experiment	6		
Cluster randomized experiment	2	**Setting**	
		Urban	8
Observed Sample Size		Suburban	5
< 50	1	Rural	3
51–100	1	Unable to determine	2
101–250	5		
251–500	6	**Classroom instructor**	
501 +	5	Teacher	17
		Researcher	1
Participants' Grade			
2–6	16	**Length**	
7–10	2	< 10 weeks	3
		10–20 weeks	3
		20–40 weeks	8
		Full School Year	4

through 6. In addition to cooperative learning, it has three core components: direct instruction in reading comprehension, story-related activities, and integrated language arts/writing. Students are invited to practice comprehension and reading skills daily in pairs and small groups. Pairs of students read to each other; predict story endings; summarize texts; write responses to questions posed by the teacher; and practice spelling, decoding, and vocabulary. Within cooperative teams, students attempt to understand the main idea of a story and work through story-specific writing activities. A Spanish version of the program is available for Grades 2 through 5 (Bilingual Cooperative and Integrated Reading and Composition). Today, the program is entitled Reading Edge or Reading Wings (depending on the grade level) and is marketed and supported by the Success for All Foundation.

Three empirical studies researched Concept Oriented Reading Instruction (CORI). CORI is a reading program (Grades 3–9) that integrates reading and science through the use of activities and science books. *CORI* is based on the rationale that when readers are fully engaged in reading, they comprehend better, use reading strategies effectively, and are motivated to read. CORI emphasizes improved comprehension by explicitly teaching students reading strategies (e.g., activating background knowledge, questioning, summarizing. CORI also focuses on increasing student reading engagement through five practices: (a) using content-area goals for a conceptual theme (e.g., survival, adaptation) during reading instruction, (b) giving students' choices and control, (c) providing hands-on activities, (d) using interesting texts for instruction, and (e) organizing opportunities for students to collaborate. Notably, both CIRC and CORI support student literacy development through the deliberate use of both collaboration and the direct instruction of reading strategies.

Mean Effects of Interventions that Utilized Within-Class Grouping

We identified 18 unique articles, reports, or conference papers that met our inclusion criteria and, within these studies, 29 unique study cohorts. Five studies (Bramlett, 1994; Calderon, Hertz-Lazarowitz, & Slavin, 1998; Guthrie et al., 1998; Guthrie, Anderson, Alao, & Rinehart, 1999; Stevens, Slavin, & Farnish, 1989) reported effect size information on multiple cohorts, and we analyzed these separately (e.g., Grade 4, 5, 6). In addition, two reports (Stevens & Durkin, 1992; Stevens, Madden, Slavin, & Farnish, 1987) included two different empirical studies conducted with distinct samples. One study (Stevens, Slavin, & Farnish, 1991) reported two comprehension outcomes for the same study cohort; in this case we calculated the effect size for each outcome, then calculated the average of the two effect sizes. The evidence base described here relies upon an observed sample of 12,286 study participants. We conducted subsequent analyses by outcome: comprehension, vocabulary, and total reading. The forest plot in Figure 1 shows the distribution of effect sizes by outcome type.

Total Reading Outcomes. Of the 29 study cohorts in this review, effect size data were reported on total reading for 16 unique cohorts. The overall weighted random effects mean was 0.16 ($p < .001$), indicating that subjects in the intervention groups had significantly higher reading achievement than comparison subjects after participating in cooperative or collaborative learning. Figure 1 shows the forest plot for the effect size distribution, using random effects methods. Notably, every effect size was positive, with effect sizes ranging from 0.06 to 0.73. The 95% confidence interval (CI) around the weighted mean [0.07, 0.25]

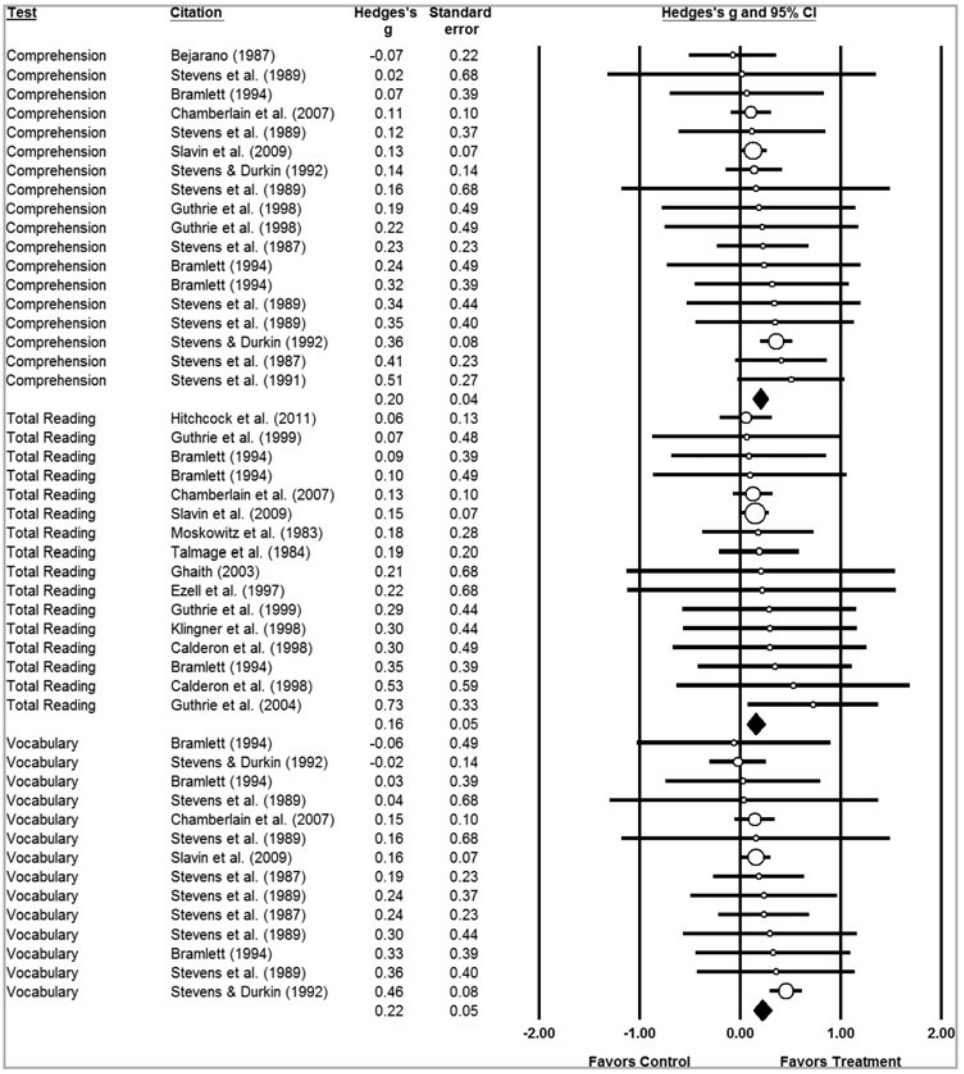

Figure 1. Forest plot of effect sizes, by outcome.

does not include zero and reveals the relative precision of the estimate of the mean random effects size of the population of studies from which these studies are presumably drawn. Table 3 summarizes this information.

Comprehension Outcomes. Of the 29 study cohorts in this review, effect size data were reported on comprehension for 18 unique cohorts. The overall weighted random effects mean was 0.20 ($p = .001$), indicating that subjects in the intervention groups had significantly higher comprehension achievement than comparison subjects after participating in a within-class grouping intervention. Figure 1 shows the forest plot for the effect size distribution, using random effects methods. The effect sizes range from –0.07 to 0.51. More

Table 3. Effects of within-class grouping, by outcome

	n	Point Estimate	SE	Variance	Lower Limit	Upper Limit	*Z* Value	*p* Value
Total reading	16	0.16	0.05	0.002	0.07	0.25	3.35	.001
Comprehension	18	0.20	0.04	0.002	0.13	0.28	5.13	<.001
Vocabulary	14	0.22	0.04	0.002	0.13	0.32	4.79	<.001

Note. These results utilize inverse variance weighting and a random effect statistical analysis.

than 90% of the effect sizes were positive. The 95% CI around this weighted mean [0.13, 0.28] does not include zero.

Vocabulary Outcomes. Of the 29 study cohorts in this review, effect size data were reported on vocabulary for 14 unique cohorts. The overall weighted random effects mean was 0.22 ($p < .001$), indicating that subjects in the intervention groups had significantly higher vocabulary achievement than comparison subjects after participating in a within-class grouping intervention. Figure 1 shows the forest plot for the effect size distribution, using random effects methods. The effect sizes range from –0.06 to 0.46. More than 85% of the effect sizes were positive. The 95% CI around this weighted mean [0.13, 0.32] does not include zero.

Homogeneity Analysis

The homogeneity analysis tests whether the observed variance is larger than expected given the subject-level sampling error within studies. We conducted tests of the homogeneity of the effect sizes using the *Q*-statistic (Hedges & Olkin, 1985) by outcome type. Analyses revealed that there was not significant variability between effect sizes to reject the null hypothesis of homogeneity. For all three outcomes, the *Q* values were not statistically significant and the *I*-squared values ranged from zero to 7.31. The *I*-squared index can be interpreted as the percentage of the total variability across studies due to heterogeneity rather than chance. Stated alternatively, the variability across effect sizes did not exceed what would be expected for subject-level sampling error alone (Lipsey & Wilson, 2001). Although heterogeneity between and among effect sizes was low, a random effects statistical model was assumed throughout the analysis. It should be noted that a random effects model reduces to a fixed effects model if the assumptions of the fixed effects model are met, which is true in this case. Because of overall homogeneity of effect sizes, there was no statistical rationale for subgroup and moderator analysis; such analyses would capitalize on chance variation rather than explore true effect size heterogeneity. Table 4 provides the results, by outcome, of the homogeneity analysis.

Table 4. Heterogeneity analyses, by outcome

	Q-value	df (*Q*)	*p* Value	*I*-squared	Tau Squared	SE	Variance	Tau
Total reading	4.71	15	0.99	0.00	0.00	0.01	0.00	0.00
Comprehension	10.24	17	0.89	0.00	0.00	0.02	0.00	0.00
Vocabulary	14.03	13	0.37	7.31	0.00	0.01	0.00	0.05

The data in Table 4 show that there was no more effect size heterogeneity than would be expected by subject level sampling error for total reading ($Q_{15} = 4.71$; $p = .99$; $I^2 = 0$), comprehension ($Q_{17} = 10.24$; $p = .89$; $I^2 = 0$), and vocabulary ($Q_{13} = 14.02$; $p = .37$; $I^2 = 7.31$). Given that the null hypothesis of homogeneity was not rejected for any the outcomes, we conducted sensitivity analyses to assess the impact of the cluster adjustments on these analyses. As previously stated, every study except for two (Chamberlain et al., 2009; Slavin, Chamberlain, Daniels, & Madden, 2009) implemented an intervention to a naturally occurring cluster of students. Based on the spreadsheet provided by McHugh (2004), the statistical procedure to account for this clustering leaves the effect size statistic relatively stable, but the effective sample sizes substantially reduced. After cluster adjustments, the median effective sample size (N) decreased by a factor of 5.7 (from 149 to 26), and the mean effective sample size (N) decreased by a factor of 3.93 (from 424 to 108). Because of these adjustments, the standard errors for the effect sizes increased by an average factor of 2.3. To determine the impact these adjustments, we reconducted the homogeneity tests with the preadjusted effect sizes and standard errors. Table 5 provides the results, by outcome, of this sensitivity analysis.

The data in Table 5 suggest that the cluster adjustments reversed our inferences about effect size heterogeneity. For example, when the total reading outcome is considered, the analysis revealed that there was significant variability to reject the null hypothesis of homogeneity ($Q_1 15 = 30.8$; $p = 0.01$). In fact, using this data, 51% of the variability across total reading effect sizes is due to sample heterogeneity rather than chance alone ($I^2 = 51.3$). For all three outcomes, moderator analyses would have been warranted without the cluster adjustments. This sensitivity analysis suggests that the cluster adjustments made an important and limiting impact on the type of analyses and inferences possible from this data. Some implications of this are discussed next (see Conclusions section).

Given the meta-analytic strategy, which utilized inverse variance weighting when synthesizing effect size data, five studies (Chamberlain et al., 2009; Hitchcock et al., 2011; Slavin et al., 2009; Stevens & Durkin, 1992; Talmage, Pascarella, & Ford, 1984) contributed the majority ($> 70\%$) of the data utilized in estimating the overall mean, weighted effect size statistics. Chamberlain et al. (2009) and Slavin, Chamberlain, Daniels, and Madden, (2009) were particularly influential because they randomly assigned students to classes, and therefore these estimates did not need to be adjusted for clustering. Overall, these studies had large effective sample sizes and their effect size estimates were therefore the most precise. To assess the impact of these studies on our meta-analytic inferences, a sensitivity analysis was conducted that removed these five studies and reanalyzed the entire data set. Overall, this sensitivity analysis had findings similar to the one just reported. For total reading, the mean weighted effect size did not change ($\mu = 0.16$; $SE = 0.07$; $p = .02$). For comprehension, the overall mean weighted effect size decreased slightly ($\mu = 0.18$; $SE = 0.08$; $p = .03$). For vocabulary, the overall mean weighted effect size decreased slightly

Table 5. Sensitivity analyses: Heterogeneity analyses before cluster adjustments, by outcome

	Q-value	*df* (*Q*)	*p* Value	*I*-squared	Tau Squared	SE	Variance	Tau
Total reading	30.80	15	0.01	51.30	0.02	0.01	0.00	0.13
Comprehension	46.05	17	0.00	63.09	0.02	0.01	0.00	0.13
Vocabulary	72.05	13	0.00	81.96	0.04	0.03	0.00	0.20

($\mu = 0.17$; $SE = 0.08$; $p = .05$). For all three outcomes, there was not significant heterogeneity to warrant moderator analysis ($I^2 = 0.0$). Overall, the inclusion of these five studies only marginally changed the magnitude of our main effect analyses and did not change the results of the heterogeneity analyses.

Publication Bias

Publication bias is potentially problematic when the results of readily available research differ from the results of all the research that has been done in an area. When this happens, readers and reviewers of that literature may end up drawing erroneous conclusions about what that body of research shows. We assessed publication bias in three ways. First, we used a weighted random effects model to analyze report type as a dichotomous moderator. For comprehension, the weighted point estimate for studies published in journals ($n = 11$) was 0.15 with a 95% CI [0.05, 0.25] that did not include zero. In contrast, unpublished reports ($n = 7$) showed a weighted point estimate of 0.30 with a 95% CI [–0.17, 0.43] that not only included zero but also completely overlapped the 95% CI for published reports. For the comprehension outcome, the difference between means was not statistically significant, $t(1) = 2.53$, $p = .94$. Similar estimates were calculated for the vocabulary and total reading outcomes, with no significant differences; $t(2) = 1.05$, 0.80 and $t(2) = 1.14$, 0.19, respectively. Overall, the weighted random effects moderator analyses did not show evidence of peer reviewed publication bias for any of the three outcomes. Table 6 presents the results of all three analyses, by outcome.

Second, we executed Duvall and Tweedie's trim and fill procedure within the CMA software. This test assumes a symmetric distribution of effect sizes around the weighted mean, imputes the presence of potentially missing studies, and recalculates the overall mean weighted effect size. As stated earlier, the random effects, mean weighted effect size

Table 6. Publication bias: Weighted regression analyses

	Total Reading		Comprehension		Vocabulary	
	Published	Unpub.	Published	Unpub.	Published	Unpub.
Effect size						
N	15	1	11	7	7	7
Point Estimate	0.17	0.06	0.15	0.30	0.16	0.26
SE	0.05	0.13	0.05	0.07	0.05	0.12
Variance	0.003	0.017	0.002	0.004	0.00	0.01
Lower limit	0.07	−0.20	0.05	0.17	0.06	0.02
Upper limit	0.27	0.32	0.25	0.43	0.27	0.49
Test of Null (two-tail)						
Z value	3.41	0.46	3.02	4.52	3.06	2.15
p value	0.001	0.64	0.003	0.00	0.002	0.03

Note. Models assumed random effects, utilized inverse variance weighting, and explored publication status as dichotomous variable.

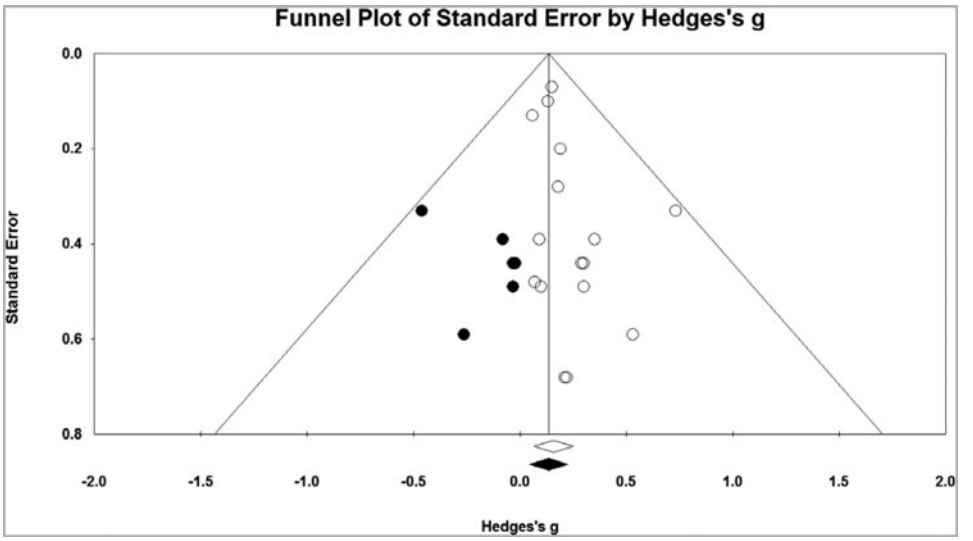

Figure 2. Funnel plot of total reading, including potentially missing studies.

estimate for total reading was 0.16, with a 95% CI that did not cross zero [0.07, 0.25]. When we adjusted this sample using the trim and fill procedure, six missing or unpublished studies were imputed. Figure 2 shows this procedure visually.

When these potentially missing studies were imputed and included in the analysis, the mean weighted effect size point estimate was reduced to 0.13. Notably, the adjusted 95% CI still did not include zero [0.05, 0.22]. This suggests that the absence of five unpublished studies would slightly reduce the overall effect size estimate; this adjustment, however, would not change the overall conclusions drawn from this sample. Table 7 includes the results for these analyses, by outcome. Using this approach to assessing publication bias, there is no reason to assume that potentially missing studies would bias our results for any of the three outcomes.

Third, we used the Egger's regression intercept test to assess publication bias. The Egger's regression intercept test uses a linear regression approach to measure funnel plot asymmetry on the natural logarithm scale of the odds ratio. The intercept provides a measure of asymmetry—the larger its deviation from zero the more pronounced the asymmetry. If there is asymmetry, with smaller studies showing effects that differ systematically from

Table 7. Publication bias: Duval and Tweedie's trim and fill method

	Total Reading		Comprehension		Vocabulary	
	Observed	Adjusted	Observed	Adjusted	Observed	Adjusted
Studies imputed		6		1		4
Point estimate	0.16	0.13	0.20	0.20	0.22	0.26
Lower limit	0.07	0.05	0.13	0.12	0.13	0.16
Upper limit	0.25	0.22	0.28	0.27	0.32	0.35
Q value	4.71	9.34	10.23	11.60	14.02	19.39

Note. Models assumed random effects and utilized inverse variance weighting.

Table 8. Publication bias: Egger's regression intercept test

	Total Reading	Comprehension	Vocabulary
Intercept	0.37	0.07	−0.26
Standard error	0.20	0.29	0.45
95% lower limit (two-tailed)	−0.05	−0.55	−1.23
95% upper limit (two-tailed)	0.79	0.70	0.72
t value	1.91	0.24	0.58
df	14	16	12
p value (one-tailed)	0.04	0.40	0.29
p value (two-tailed)	0.08	0.81	0.58

larger studies, the regression line will not run through the origin. Table 8 shows results from these analyses, by outcome type.

For total reading, the Egger's regression method for the 14 effect sizes produced an intercept (β_0) of 0.37 and a 95% CI that marginally included zero [–0.05, 0.79]. For total reading, the Egger's regression test suggested that there is some evidence of publication bias, $t(14) = 1.91$, $p = .08$, indicating that smaller studies tended to report smaller effect sizes and studies of negative and/or lower effects may be missing. For the vocabulary and comprehension outcomes, the Egger's regression interception technique did not support the conclusion of publication bias.

Collectively, these three tests suggest that (a) peer-reviewed publication is not a statistically significant moderator of effects size, that (b) potentially absent or unpublished studies might reduce the overall mean effect size but would not affect our overall conclusions, and that (c) the current effect size distribution is symmetrical enough to continue without serious concern about publication bias. It is important to note that these tests do not disprove the possibility of publication bias; they only suggest that this sample of effect sizes is not consistent with the presence of publication bias.

CONCLUSIONS

The objective of this review was to gather, summarize, and synthesize the empirical findings on cooperative and collaborative activity structures to support literacy achievement. Searching more than 30 years of research, we located 18 studies with 29 cohorts and found a positive effect overall. When teachers organized student learning using cooperative and collaborative grouping, their students showed evidence of enhanced achievement in vocabulary, comprehension, and total reading. Nearly 90% of the effect size values were positive. Considering all three outcomes, the overall weighted mean effect size ranged from +0.16 to +0.22, and all of these were statistically significant. These conclusions are very similar to the findings reported by Lou et al. (1996), who found who found an overall mean weighted effect size of +0.17 in favor of small-group instruction when examining mathematics and other outcomes. These conclusions are also consistent with the findings of more recent

reviews (Slavin et al., 2008; Slavin, Lake, Chambers, Cheung, & Davis, 2009) that argued that cooperative learning is a core component of the most effective reading programs.

Although these interventions, as a whole, positively supported student literacy achievement, it was impossible to disentangle the effects of cooperative/collaborative learning from overall intervention effects. In this body of research, cooperative or collaborative learning was never investigated in isolation; every intervention utilized cooperative/collaborative learning along with other instructional components. Typically, interventions combined cooperative/collaborative learning with explicit reading strategy instruction (e.g., summarize, predict, clarify, question), and, in many cases, with supplementary curriculum tasks and texts. While this is a restriction on meta-analytic inferences, this boundary illustrates—to a large degree—the complexity and multidimensionality of modern literacy programs, which explicitly and implicitly utilize a diverse collection of tasks, texts, tools, and modes of argumentation (Lehrer & Schauble, 2000).

This review has important implications for the research community. Although within-classroom "ability" grouping has been criticized for widening the achievement gap and providing inferior instruction for lower "ability" students in general, this meta-analysis shows that the cooperative and collaborative activity structures used in CIRC, CORI, and other literacy programs are effective at improving student literacy achievement. The one study (Bramlett, 1994) that specifically examined the differential influence that cooperative learning had on students of low, medium, and high "ability" found the *most* positive effects for students who initially had the lowest literacy achievement. By design, cooperative and collaborative activity structures deliberately create and construct heterogeneous and "mixed ability" groups.

In addition, this review finds some important patterns and trends among the research. Although guided reading (Fountas & Pinnell, 1996) is currently a highly touted and widely used instructional practice (e.g., Avalos, Plasencia, Chavez, & Rascón, 2007; Frey & Fisher, 2010), we could not find a single study on guided reading employed an experimental or quasi-experimental design. This is an important area for future research. In addition, although we did not strategically set out to review studies that investigated how social media and technology (e.g., Facebook, wikis, blogs) can support student literacy development, this would be an important topic for future reviews, particularly because many forms of media and technology implicitly utilize distributed (e.g., non-face-to-face) cooperation and collaboration activity structures.

This study also has important implications for meta-analyses. As noted in the results section, the cluster adjustments (Hedges, 2004a, 2004b, 2007) reduced our mean sample size by a factor of 5.7, thus expanding the standard errors by a factor of 2.3. The sensitivity analyses conducted indicated that the inferences drawn from this data may have been different if these cluster adjustments were not made. Without cluster adjustments, apparent effect size heterogeneity would have warranted moderator analyses. Given that all but two studies were conducted with previously in-tact classrooms, however, we believe that these cluster adjustments were necessary and important for future meta-analysts to consider. Naturally occurring classrooms cannot be viewed as statistically independent and analyzed as if they were a simple random sample. We believe that conducting moderator analyses would have capitalized on chance rather than true variation and led to unwarranted inferences.

Last, and most important, we hope that this review will help teachers, principals, and policymakers as they make decisions regarding the organization of classroom literacy instruction. Teachers have a wide variety of pedagogical choices and attempt to support the learning and development of their students with effective, research-based practices.

Likewise, principals and districts want to support teachers' learning and development of such practices. Although teachers may report the regular use of grouping, the best observational estimates that we have today (Cordray et al., 2011) suggest that teachers only group students for literacy instruction between 30 and 40% of the time. Although the limited number of eligible studies prevents us from making claims about specific forms of grouping that are more effective and the circumstances under which programs have more impact, our findings strongly suggest that cooperative and collaborative grouping is a core component of effective literacy interventions, particularly at the elementary level.

ACKNOWLEDGMENTS

The work of Kelly Puzio and Glenn T. Colby was supported by Vanderbilt University's Experimental Education Research Training (ExpERT) grant (David S. Cordray, Director; IES Grant No. R305B040110). The opinions expressed are those of the authors and do not represent the views of the U. S. Department of Education.

REFERENCES

Note: References marked with an asterisk (*) indicate studies included in the meta-analysis.

Allington, R. L. (1983). The reading instruction provided readers of differing reading abilities. *The Elementary School Journal*, *83*, 548–559.

Austin, M. C., & Morrison, C. (1963). *The first R: The Harvard report on reading in elementary schools*. New York, NY: Macmillan.

Avalos, M. A., Plasencia, A., Chavez, C., & Rascón, J. (2007). Modified guided reading: Gateway to English as a second language and literacy learning. *The Reading Teacher*, *61*, 318–329.

Baker, L., Dreher, J., & Guthrie, J. (Eds.). (2000). *Engaging young readers: Promoting achievement and motivation*. New York, NY: Guilford.

Baumann, J. F., Hoffman, J. V., Duffy-Hester, A. M., & Moon Ro, J. (2000). The first R yesterday and today: U.S. elementary reading instruction practices reported by teachers and administrators. *Reading Research Quarterly*, *35*, 338–377.

*Bejarano, Y. (1987). A cooperative small group methodology in the language classroom. *TESOL Quarterly*, *21*, 483–504.

Borko, H., & Eisenhart, M. (1989). Reading groups as literacy communities. In D. Bloome (Ed.), *Communities and Literacy* (pp. 107–132). Norwood, NJ: ABLEX.

Braddock, J. H., & Slavin, R. E. (1992). Why ability grouping must end: Achieving excellence and equity in American education. *Journal of Intergroup Relations*, *20*(1), 51–64.

*Bramlett, R. (1994). Implementing cooperative learning: A field study evaluating issues for school-based consultants. *Journal of School Psychology*, *32*, 67–84.

*Calderon, M., Hertz-Lazarowitz, R., Slavin, R. (1998). Effects of bilingual cooperative integrated reading and composition on students making the transition from Spanish to English reading. *The Elementary School Journal*, *99*, 153–165.

*Chamberlain, A., Daniels, C., Madden, N. A., & Slavin, R. E. (2009). A randomized evaluation of the Success for All Middle School reading program. *Middle Grades Research Journal*, *2*(1), 1–21.

Chorzempa, B. F., & Graham, S. (2006). Primary-grade teachers' use of within-class ability grouping in reading. *Journal of Educational Psychology*, *98*, 529–541.

Cordray, D., Pion, G., Dawson, M., Brandt, C., & Molefe, A., (2011). *The Impact of the Measures of Academic Progress (MAP) on Differentiated Instruction and Student Achievement, Year 2 Report* (REL Midwest at Learning Point Associates, ED-06-CO-0019).

Elbaum B., Vaughn, S., Hughes, M., & Moody, S. W. (1999). Grouping practices and reading outcomes for students with disabilities. *Exceptional Children*, *65*, 399–415.

Englert, C. S., & Mariage, T. V. (1991). Making students partners in the comprehension process: Organizing the reading "posse." *Learning Disability Quarterly*, *14*, 123–138.

Evans, M. M. (1942). *The effect of variable grouping on reading achievement* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

*Ezell, H. (1997). Use of peer-assisted procedure to teach QAR reading comprehension strategies to third-grade children. *Education and Treatment of Children*, *15*, 205–227.

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., & Shelley, B. E. (2010). *Highlights from PISA 2009: Performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international context* (NCES 2011–004). Washington, DC: U.S. Government Printing Office.

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann Press.

Frey, N., & Fisher, D. (2010). Identifying instructional moves during guided learning. *The Reading Teacher*, *64*(2), 84–95.

Gamoran, A. (1992). Synthesis of research/Is ability grouping equitable? *Educational Leadership*, *50*(2), 11–17.

*Ghaith, G. (2003). Effects of the learning together model of cooperative learning on English as a Foreign language reading achievement, academic self-esteem, and feelings of school alienation. *Bilingual Research Journal*, *27*, 451–474.

*Guthrie, J., Anderson, E., Alao, S., & Rinehart, J. (1999). Influences of concept-oriented reading instruction on strategy use and conceptual learning from text. *The Elementary School Journal*, *99*, 343–366.

*Guthrie, J., Van Meter, P., Hancock, G. R., Alao, S., Anderson, E., & McCann, A. (1998). Does concept-oriented reading instruction increase strategy use and conceptual learning from text? *Journal of Educational Psychology*, *90*, 261–278.

*Guthrie, J., Wigfield, A., Barbosa, P., Perencevich, K., Taboada, A., Davis, M. , ... Tonks, S. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology*, *96*, 403–423.

Hedges, L. (2004a). Correcting significance tests for clustering. Unpublished manuscript.

Hedges, L. (2004b). *Effect sizes in multi-site designs using assignment by cluster*. Unpublished manuscript.

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

Hedges, L., & Hedberg, E. (2007). Intraclass correlation values for planning group-randomized trials in education. *Journal of Educational Evaluation and Policy Analysis*, *29*, 60–87.

Hedges, L., & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London, UK: Academic Press.

Hiebert, E. H. (1983). An examination of ability grouping for reading instruction. *Reading Research Quarterly*, *18*, 231–255.

*Hitchcock, J., Dimino, J., Kurki, A., Wilkins, C., & Gersten, R. (2011). *The impact of collaborative strategic reading on the reading comprehension of Grade 5 students in linguistically diverse schools* (Report NCEE 2011-4001). Department of Education. Washington, DC: Government Printing Office.

Jenkins, J. R., Jewell, M., Leicester, N., & O'Conner, R. E. (1994). Accommodations for individual differences without classroom ability groups: An experiment in school restructuring. *Exceptional Children*, *60*, 344–358.

Johnson, D. W. & Johnson, R. W. (1986). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Books.

*Klingner, J. K., Vaughn, S., & Schumm, J. S. (1998). Collaborative strategic reading during social studies in heterogeneous fourth-grade classrooms. *The Elementary School Journal*, *99*, 3–22.

Kulik, J. A. (1992). *An analysis of research on ability grouping: Historical and contemporary perspectives (*Research-*based Decision-Making Series)*. Storrs, CT: University of Connecticut, National Research Center on the Gifted and Talented. (ERIC Document Reproduction Service No. ED 350 777)

Kulik, J. A., & Kulik, C.-L. C. (1987). Effects of ability grouping on student achievement. *Equity and Excellence*, *23*, 22–30.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In *Advances in instructional psychology* (Vol. 5, pp. 101–159). Mahwah, NJ: Erlbaum.

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

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. *Journal of Educational Research*, *94*, 101–112.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, *66*, 423–458.

McHugh, C. M. (2004). *Calculations for correcting test statistics and standard errors for clustering* [Excel spreadsheet].

Moskowitz, J. M., Malvin, J. H., Schaeffer, G. A., & Schaps, E. (1983). Evaluation of a cooperative learning strategy. *American Educational Research Journal*, *20*, 687–696.

National Center for Education Statistics. (2009). *The Nation's Report Card: Reading 2009* (NCES 2010–458). Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Oakes, J. (1985). *Keeping track: How schools structure inequality.* New Haven, CT: Yale University Press.

OECD. (2010). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science (Volume I)*. http://dx.doi.org/10.1787/9789264091450-en

Oxford, R. L. (1997). Cooperative learning, collaborative learning, and interaction: Three communicative strands in the language classroom. *Modern Language Journal*, *81*, 443–456.

Peterson, J. M. (1989). Tracking students by their supposed abilities can derail learning. *American School Board Journal*, *176*(5), 38–46.

Rapp, J. C. (1991). *The effect of cooperative learning on selected student variables* (Doctor of education dissertation, Washington State University, 1991). Dissertation Abstracts International, 52(10–A), 3516.

Reutzel, D. R., & Cooter, R. B. (1991). Organizing for effective instruction: The reading workshop. *The Reading Teacher*, *44*, 548–554.

Rosenbaum, J. (1976). *Making inequality: The hidden curriculum of high school tracking*. New York, NY: Wiley.

Shields, J. M. (1927). Teacher reading through ability grouping. *The Journal of Educational Method*, *7*, 7–9.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*, 293–336.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*, 417–499.

*Slavin, R., Chamberlain, A., Daniels, C., & Madden, N. A. (2009). The Reading Edge: A randomized evaluation of a middle school cooperative reading program. *Effective Education*, *1*, 13–26.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, *43*, 290–322.

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.

*Stevens, R. & Durkin, S. (1992, September). *Using student teams reading and student team writing in middle schools. Two evaluations* (Report No. 36). Baltimore, MD: John Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

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

*Stevens, R., Slavin, R., & Farnish, A. (1989, November). *A cooperative learning approach to elementary reading and writing instruction: long-term effects* (Center for Research on Elementary and Middle Schools. Report No. 42).

*Stevens, R. J., Slavin, R. E., & Farnish, A. M. (1991). The effects of cooperative learning and instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology*, *83*, 8–16.

*Talmadge, H., Pascarella, E. T., & Ford, S. (1984). The influence of cooperative learning strategies on teacher practices, student perceptions of the learning environment, and academic achievement. *American Educational Research Journal*, *21*, 163–179.

Top, B. L., & Osguthorpe, R. T. (1987). Reverse-role tutoring: The effects of handicapped students tutoring regular class students. *Elementary School Journal*, *87*, 413–423.

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Weinstein, R. (1976). Reading group membership in the first grade: Teacher behaviors and pupil experience over time. *Journal of Educational Psychology*, *68*, 103–116.

Wheelock, A. (1994). *Alternatives to tracking and ability grouping*. New York, NY: R&L Education.

APPENDIX

TABLE A1. Summary of Included Studies

					Comprehension	Vocabulary	Total Reading	
Citation	Program	Cohort	N	Adj. N	Hedges's *g*	Hedges's *g*	Hedges's *g*	Test
Bejarano (1987)	STAD	Grade 7	467	82	−0.07 (0.22)†			RD
Bramlett (1994)	CIRC	Lowest 33%	149	26	0.32 (0.39)	0.33 (0.39)	0.35 (0.39)	CAT
	CIRC	Middle 33%	151	26	0.07 (0.39)	0.03 (0.39)	0.09 (0.39)	CAT
	CIRC	Upper 34%	92	17	0.24 (0.49)	−0.06 (0.49)	0.10 (0.49)	CAT
Calderon et al. (1998)	BCIRC	Grade 2	93	17			0.30 (0.49)	STAAS
	BCIRC	Grade 3	85	13			0.53 (0.59)	ENAP
Chamberlain et al. (2009)	RE	Grade 6	405	405+	0.11 (0.10)	0.15 (0.10)	0.13 (0.10)	GM
Ezell et al. (1997)	QAR	Grade 4	48	9			0.22 (0.68)†	CAT
Ghaith (2003)	LT	Grade 9	56	9			0.21 (0.68)	RD
Guthrie et al. (1998)	CORI	Grade 3	90	17	0.19 (0.49)†			RD
	CORI	Grade 5	82	17	0.22 (0.49)†			RD
Guthrie et al. (1999)	CORI	Grade 3	120	21			0.29 (0.44)	CTBS
	CORI	Grade 5	101	17			0.07 (0.48)†	MAT
Guthrie et al. (2004)	CORI	Grade 3	267	47			0.73* (0.33)	GM
Hitchcock et al. (2011)	CSR	Grade 5	1355	233			0.06 (0.13)	GR
Klingner et al. (1998)	CS	Grade 4	141	22			0.30 (0.44)†	GM

(Continued on next page)

TABLE A1. Summary of Included Studies *(Continued)*

					Comprehension	Vocabulary	Total Reading	
Citation	Program	Cohort	N	Adj. N	Hedges's *g*	Hedges's *g*	Hedges's *g*	Test
Moskowitz et al. (1983)	CL	Grade 5, 6	239	53			0.18 (0.28)	SAT
Slavin et al. (2009)	RE	Grade 6	788	788+	0.13 (0.07)	0.16* (0.07)	0.15* (0.07)	GM
Stevens et al. (1987)	CIRC	Grade 3, 4	461	78	0.23 (0.23)†	0.19 (0.23)†		CAT
	CIRC	Grade 3, 4	447	78	0.41 (0.23)†	0.24 (0.23)†		CAT
Stevens et al. (1989)	CIRC	Grade 2	135	22	0.34 (0.44)	0.30 (0.44)		CAT
	CIRC	Grade 3	53	9	0.16 (0.68)	0.16 (0.68)		CAT
	CIRC	Grade 4	135	25	0.35 (0.40)	0.36 (0.40)		CAT
	CIRC	Grade 5	157	30	0.12 (0.37)	0.24 (0.37)		CAT
	CIRC	Grade 6	49	9	0.02 (0.68)	0.04 (0.68)		CAT
Stevens et al. (1991)	CIRC	Grade 3, 4	320	56	0.51 (0.27)			RD
Stevens & Durkin (1992)	STR/W	Grade 6	1223	211	0.14 (0.14)†	−0.02 (0.14)†		CAT
	STR/W	Grade 6, 7, 8	3986	689	0.36*** (0.08)	0.46*** (0.08)		CAT
Talmage et al. (1984)	CL	Grade 3, 4	591	103			0.19 (0.20)	SRA

Note. Standard errors are in parentheses. RD = researcher designed; ITBS = Illinois Test of Basic Skills; CAT = California Achievement Test; GM = Gates-MacGinitie Reading Test; STAAS = Spanish Texas Assessment of Academic Skills; ENAP = English Norm-Referenced Assessment Program; MAT = Metropolitan Achievement Test; GR = Group Reading Assessment and Diagnostic Evaluation; CTBS = California Test of Basic Skills; SRA = Science Research Associates; SAT = Stanford Achievement Test; CIRC-NW = CIRC (no writing); CS = Collaborative Strategic; LT = Learning Together, CL = Cooperative Learning, CSR = Collaborative Strategic Reading; Student Teams Reading & Writing; RE = Reading Edge.

†In cases where posttest results were reported without using pretest scores as covariates, we subtracted pretest differences.

+Only two studies randomly assigned students to classes and were therefore not adjusted for clustering.

*$p < .05$. **$p < .01$. ***$p < .001$.