Campbell Systematic Reviews
2006:7
First published: 1 May, 2006
Last updated: 1 May, 2006

The Effectiveness of Volunteer Tutoring Programs: A Systematic Review

Gary Ritter, George Denny, Ginger Albin, Joshua Barnett, Virginia Blankenship

THE CAMPBELL COLLABORATION

Colophon

Title	The effectiveness of volunteer tutoring programs: A systematic review
Institution	The Campbell Collaboration
Authors	Ritter, Gary Denny, George Albin, Ginger Barnett, Joshua Blankenship, Virginia
DOI	10.4073/csr.2006.7
No. of pages	63
Last updated	1 May, 2006
Citation	Ritter G, Denny G, Albin G, Barnett J, Blankenship V. The effectiveness of volunteer tutoring programs: A systematic review. Campbell Systematic Reviews 2006:7 DOI: 10.4073/csr.2006.7

Keywords	
Contributions	Gary Ritter (GR) wrote the review. George Denny (GD) led the meta-analysis, with assistance from GR. Ginger Albin (GA) and Joshua Barnett (JB) worked with GR in developing the search strategy and then implemented the search strategy. GD, with help from GA and JB, extracted data from all of the fully coded articles. Virginia Blankenship worked with GR in initial article search and in developing protocol.
Support/Funding	University of Arkansas, USA Smith Richardson Foundation, USA (grant to the University of Pennsylvania) Smith Richardson Foundation, USA (grant to the University of Missouri)
Potential Conflicts of Interest	Gary Ritter was not involved in the decision whether to include the Ritter 2000 study.

Corresponding author	Gary W. Ritter Education and Public Policy University of Arkansas 207 Graduate Education Building Fayetteville, AR 72701 USA Telephone: +1 479 575 4971 E-mail: garyr@uark.edu

Campbell Systematic Reviews

The Campbell Collaboration (C2) was founded on the principle that systematic reviews on the effects of interventions will inform and help improve policy and services. C2 offers editorial and methodological support to review authors throughout the process of producing a systematic review. A number of C2's editors, librarians, methodologists and external peer-reviewers contribute.

The Campbell Collaboration
P.O. Box 7004 St. Olavs plass
0130 Oslo, Norway
www.campbellcollaboration.org

Cover sheet

Title
The effectiveness of volunteer tutoring programs: A systematic review

Reviewers
Ritter G, Denny G, Albin G, Barnett J, Blankenship V

Dates
Date edited:
Date of last substantive update:
Date of last minor update:
Date next stage expected:
Protocol first published
Review first published:

Contact Reviewer
Dr. Gary W. Ritter
Associate Professor, Education and Public Policy
University of Arkansas
207 Graduate Education Building
Fayetteville, AR 72701
Phone: (479) 575-4971
Fax: (479) 575-3196
Email: garyr@uark.edu
URL: http://policy.uark.edu/ritter

Contribution of Reviewers
GR (lead reviewer) wrote the review. GD led the meta-analysis, with assistance from GR. GA and JB worked with GR in developing the search strategy and then implemented the search strategy. GD, with help from GA and JB, extracted data from all of the fully coded articles. VB worked with GR in initial article search and in developing protocol.

Internal Sources of Support
University of Arkansas

External Sources of Support
The completion of this review during 2005 was supported through a grant from the Campbell Collaboration—Education Review Group. Earlier support for this project was provided by the Smith Richardson Foundation under grants to the University of Pennsylvania and the University of Missouri.

The draft protocol for the pilot review was completed with the help of many colleagues, including participants in the University of Pennsylvania's Graduate School of Education seminar on conducting systematic reviews and meta-analysis, led by Dr. Rebecca Maynard.

The initial investigation into the field of volunteer tutoring was supported under the "Evaluation of the West Philadelphia Tutoring Project" conducted by faculty and staff at the University of Pennsylvania from 1998-2000.

The systematic review is now being conducted at and funded by the University of Arkansas. The results of this study do not necessarily represent the official opinion or policy of the Smith Richardson Foundation, the University of Pennsylvania, or the University of Arkansas.

Acknowledgments

Thanks to Rebecca Maynard for oversight and guidance in getting this review started and throughout. Thanks to Geraldine Macdonald and Jeff Valentine for helpful comments and suggestions on this review. This review benefited from technical advice from Mark Lipsey, Chad Nye, Herb Turner, and Robert Boruch.

Potential Conflict of Interest

Gary Ritter was not involved in the decision whether to include the Ritter 2000 study.

Dates

Date review re-formatted:
Date new studies sought but none found:
Date new studies found but not yet included/excluded:
Date reviewers' conclusions section amended:
Date comment/criticism added:
Date response to comment/criticisms added:

Text of Review

Synopsis

Based on increasing demands for accountability in schools, school administrators are looking for ways to increase the academic performance of students including tutoring. Currently a variety of tutoring approaches are used including volunteer tutoring, peer tutoring, cross-age tutoring, and one-on-one tutoring. The evidence base related to volunteer tutoring has grown in recent years; consequently, this review is based upon only those evaluations of volunteer programs in which tutees were randomly assigned to a treatment or control group. We found 21 studies (with 27 different study cohorts in those studies because several studies provided separate reporting on multiple cohorts) reporting on randomized field trials to guide us in assessing the effectiveness of volunteer tutoring programs.

Abstract

Background

Volunteer tutoring programs are intended to improve student performance, provide mentorship, and improve student self-esteem, as well as behavior. Despite the best of intention and effort, schools are not certain which volunteer tutoring programs are most effective. Therefore, we contend that a rigorous analysis of the extant literature regarding volunteer tutoring programs can provide schools with information about the most effective types of tutoring programs.

Objectives

To assess the effectiveness of volunteer tutoring programs for improving the academic skills of student enrolled in grades K-8 in the USA, and to further investigate for whom and under what conditions tutoring can be effective.

Search Strategy

A range of electronic social science databases were searched including Academic Search Premier; Primary Search; Professional Development Collection; Middle Search Plus; Psychology and Behavioral Sciences Collection; PsycINFO; Sociological Collection, ERIC (Education Resources Information Center), and Proquest Digital Dissertations.

Selection Criteria

Only randomized field trials (in which a treatment was compared to a control group that did not receive the treatment) published from January 1985 to August 2005 were included in our review. We only included studies that included academic impacts, although other impacts could be part of the study (e.g. behavioral, emotional). We also only included programs that were used for students in grades K – 8, and programs where adult, non-professional (volunteer) tutors were used.

Data Collection & Analysis

Details on the methods and procedures (i.e., design, analysis, outcome measures), the intervention (i.e., duration, setting), and the subject samples (i.e., age, gender, ethnicity) were coded for each study to allow for analysis of outcomes for different types of volunteer tutoring programs and for different study characteristics. After we identified all outcomes measured in the included studies, we selected the appropriate effect size metric for the meta-analysis. We used Hedges' unbiased estimate *g* of the standardized mean difference effect size statistic (the difference between the treatment and control group means on an outcome variable divided by their pooled standard deviations) for each outcome measure. When means and standard deviations were not reported, we attempted to estimate the effect sizes using the procedures recommended by Wilson and Lipsey (2001). When studies had multiple measures of an outcome, we computed a pooled mean of effect size.

Main Results

The results of the review were based on the data from 1,676 study participants in 28 study cohorts in 21 research articles or reports. The analysis of these studies – most of which included relatively small samples – showed that volunteer tutoring programs can positively influence language and reading outcomes for students. We began by examining the overall effects of volunteer tutoring on student reading outcome measures. Twenty-five studies assessed reading measures of one type or another. The average effect of volunteer tutoring programs on reading outcomes for elementary students is 0.23. After removing one outlier study which disproportionately influenced by the overall result, we found an average effect size of 0.30. We also found several significant results in the meta-analyses of specific academic domains. The outcomes where volunteer tutoring programs made a significant difference were Reading Global (effect size = 0.26), Letters and Words (effect size = 0.41), Oral Fluency (effect size = 0.30), and Writing (effect size = 0.45). We found positive, but not significant, effects of volunteer tutoring on Reading – Comprehension and Mathematics.

When we analyzed the reading outcomes separately by study characteristics, we found no significant difference in effect size by tutor type, grade level, or program focus. Highly structured tutoring programs had a significantly greater effect on global reading outcomes than programs with low structure, but not on the other outcome types. The difference in effect sizes between studies published in journals and non-published studies was not statistically significant. Other tests of publication bias also suggested the included studies were an unbiased sample.

Reviewers' Conclusions

This review does not suggest that there are any particular volunteer tutoring models that should be recommended for immediate adoption for schools and districts across the country. Rather, we can conclude from the analysis that these programs can positively influence important reading and language sub-skills for young students. The results are substantial – approximately one-third standard deviation. In the end, the results of this analysis should serve as one important piece of evidence used by policymakers and educators who are deciding whether to employ volunteer tutoring as a strategy to improve academic skills for young students. As educators across the country work to meet adequate yearly progress goals in state accountability systems, and as they seek affordable ways to offer additional services to students at risk of not meeting annual academic goals, it would be worthwhile to consider structured, reading-focused volunteer tutoring programs as strategies to improve reading and language skills.

1. Background

In many cultures, the oldest form of teaching was provided to children by in-home tutors or private instructors (Shanahan, 1998). Tutoring remains a popular form of instruction worldwide, and the effectiveness of tutoring as a pedagogical method has been documented extensively in various strands of the educational literature (see, for example, Cohen, Kulik, & Kulik, 1982; Fashola, 2001; Wasik & Slavin, 1993).

During the 1970s, U.S. schools began relying more on peer tutoring (also known as student-to-student or cross-age tutoring) as a way to efficiently use scarce resources in a period of teacher shortages (Rekrut, 1994). The next two decades also witnessed an increased interest in tutoring programs staffed by adult volunteers for a variety of reasons: (1) increased public concern with the quality of education after the National Commission on Excellence in Education's release of "A Nation at Risk" in 1983; (2) rising interest in community service in the 1990s, and (3) the encouraging results from effective yet costly programs that employ professional tutors. By 1987, the National Research Council estimated that there were over one million volunteer tutors who donated an average of four hours per week in the nation's public schools. The survey found that three-fourths of public elementary schools in the United States reported the involvement of volunteers, with schools having an average of 24 volunteers (Michael, 1990).

More recently, the growth of these university partnerships has accelerated due to the America Reads Challenge, a nationwide tutoring initiative launched in 1997 by President Clinton. At that time, nearly 800 universities and colleges throughout the nation had already pledged to commit work-study slots for college students to serve as tutors for elementary school children (White House, 1997). By 1999, nearly 1,200 colleges and universities committed to placing work-study students as tutors in public schools. President Clinton's 1999 budget proposal also included $140 million to establish programs matching university-based mentors with students in schools that had very high dropout rates and high concentrations of poor students. As a result of the America Reads Challenge, state leaders have become increasingly interested in providing tutoring programs for elementary school children, and numerous local tutoring initiatives are now receiving increased support (U.S. Department of Education 1996; 1997).

Despite the increased interest in and support for tutoring programs over the past few decades, the expansion of programs that use non-professional, adult volunteer tutors has yet to be matched by a supporting research base. In the 1970s and early 1980s, most tutoring research in the United States focused on the impacts of peer or cross-age tutoring (Shanahan, 1998). In 1982, Cohen, Kulik, and Kulik published a well-known and often-cited review of this early body of tutoring research in the *American Educational Research Journal*. The authors noted that hundreds of reports on tutoring had been written by teachers and researchers, some based on scientifically-sound, experimental design evaluations, and others more informal and subjective. According to the authors, the four major reviews of the research up to 1982 used "relatively informal narrative and box score techniques" (p. 234), and each concluded that peer tutoring can help improve the academic success of young students.[1]

During the 1980s and 1990s, researchers began focusing more on specialized interventions aimed at improving the academic achievement of the lowest-achieving children, most notably Reading Recovery and Success for All (Shanahan, 1998; Wasik & Slavin, 1993). Both programs include

[1] The four reviews are: Devin-Sheehan, Feldman, & Allen, 1976; Ellson, 1976; Fitz-Gibbon, 1977; and Rosenshine & Furst, 1969.

one-on-one tutoring by professional tutors and are perceived as effective by many in the research community (Wasik, 1998). However, because the expense of employing professional tutors limits the number of children who can be served by these interventions, several programs have been created in recent years that use adult volunteers or para-professionals as tutors. Some of these programs are responses to the accountability measures established by the No Child Left Behind Act of 2001, which encourage education administrators to implement new programs aimed at increasing student performance. As administrators encourage the use of tutoring programs in their schools, they may be doing so without solid evidence of what types of impacts specific tutoring programs can realistically produce. To begin understanding the effects of tutoring programs, we examined the existing literature for reviews.

Previous work has identified four major reviews of research on the impact of volunteer tutoring programs on student outcomes: Topping and Hill (1995), Wasik (1998), Shanahan (1998), and Elbaum et al. (2000).

In 1995, Topping and Hill provide interesting background on the research related to volunteer tutoring in a chapter contributed to the book, Students as tutors and mentors (Goodlad, 1995). The chapter presents a review of the evaluation research around the world related to the effectiveness of college students as tutors for schoolchildren.

Wasik (1998) reviewed studies of 17 programs that used volunteer tutors to help improve students' reading abilities. While the evidence suggested that volunteers can indeed help many children improve their reading skills, the results varied considerably across programs. Furthermore, only two of the 17 programs reviewed compared students' achievement with that of a control group, which makes the results less certain. Wasik's findings are consistent with those of the National Research Council's 1998 report, Preventing Reading Difficulties in Young Children, which concluded that there is "[no] evidence confirming that [volunteers] are able to deal effectively with children who have serious reading problems" (p. 238). Nevertheless, it is likely that there are other evaluations not included in either of these non-systematic reviews, which could have altered the findings.

In contrast, Shanahan's (1998) review of the research on volunteer tutoring found that, despite many limitations, these programs can indeed be effective in improving student achievement. However, Shanahan offers little detail about the methodology used in his review, and, as with the reviews listed above, the results of several important studies may well be excluded.

Most recently, Elbaum et al. (2000) reported on a meta-analysis in the Journal of Educational Psychology focused on the effectiveness of one-to-one tutoring programs for improving reading ability in elementary students at risk for reading failure. The authors reviewed 29 studies involving 42 samples of students between 1975 and 1998 and found that trained tutors can help students improve in reading skills. This review is helpful, but only focused on reading as a subject matter and included all types of tutoring, including that done by trained professions.

The limited and often-conflicting evidence on the effectiveness of volunteer tutoring programs leaves us with many questions. First and foremost, is there good evidence to encourage policymakers and school leaders to continue to pursue volunteer tutoring as a possible strategy for improving the academic skills of young students? How do differences in participants (e.g., age, gender) impact the effectiveness of the programs? And how do differences in tutoring programs (e.g., program structure, program focus, types of tutors) affect the results?

2. Objectives

The objective of this systematic review was to gather, summarize, and integrate the empirical research on volunteer tutoring programs in order to help policymakers, educators, parents, and other stakeholders understand whether this type of intervention might be an effective tool for improving academic skills for elementary students.

3. Methods of the Review

Criteria for Considering Studies for Review

Types of Studies

Only randomized field trials were included in the review. Quasi-experimental studies that employ treatment and control groups matched on pre-tests of key outcome variables were not included in this review. Pretest-posttest studies, or those in which a treatment group is compared to another treatment group, were not included. Studies published before 1985 were not included. Only English-language studies of programs conducted in the United States were considered, due to the limited resources for this review. Furthermore, we excluded studies of programs that were especially designed to address the needs of students with limited English proficiency (LEP), because such specialized programs are not representative of most volunteer tutoring programs for elementary and middle school students.

Types of Participants

Only studies of programs involving adult, non-professional tutors were included. Although these tutors were almost always referred to as "volunteers" in the literature, those programs that pay a small stipend to tutors (such as undergraduate tutors who are tutoring as part of a work-study program) were also included. Several of the programs in this area are those that train parents to tutor their own students; such programs are included in this review, and are coded for separate examination in the subgroup analysis. In terms of the tutees, only studies of programs that serve students in grades K-8 (elementary and middle school) were considered, since this is the population typically served by volunteer tutoring programs, and because such programs are fundamentally different than those provided to high school students.

Types of Interventions

The interventions featured regular tutoring sessions with an academic focus for at least one month in duration. The duration restriction was included due to a belief that programs lasting only a few days were qualitatively different than longer programs with sustained exposure. With regard to intervention focus, we did include interventions with other components in addition to an academic focus (such as behavior modification); however, the evaluation had to focus on at least one academic outcome measure.

Types of Outcome Measures

The original intent of the review was to consider all outcome measures related to student achievement, including distal outcomes (ones the program are actually intended to influence, such as school grades or standardized achievement measures), as well as proximal outcomes (intermediate measures that might be influenced by tutoring and then might lead to improved outcomes in the future, such as student attendance rates). However, the review yielded very few studies that analyzed school grades or attendance rates; rather, most studies focused on various

standardized assessments of math and reading skills, or "authentic" measures of reading and writing skills. As a result, our review focused on these outcomes.

Many of the included studies, particularly those with a reading focus, examined several outcomes and it was the task of the reviewers to categorize those outcomes within our classifications of outcome measures. We organized the outcomes into six broad categories based on different concepts described within the retained articles. Listed below are the six classifications and the outcome measures that fit within each class. (Table 11 in Section 12 presents each class of outcome measures along with the individual outcomes from each study that are included within the broad class.)

Reading Global. In this category, the reviewers included results from overall batteries on standardized reading achievement tests. The achievement tests in this classification include the Gates MacGinitie Reading Test, the Wide Range Achievement Test reading section, the Comprehensive Test of Basic Skills reading section, and the reading battery on the Stanford Achievement Test.

Reading Letters and Words. Many of the reading-focused studies examined multiple outcomes related to reading sub-skills that the reviewers define here as being in the "Letters and Words" category. The types of measures that are included in this class are those that focus on decoding of words and knowledge of words. The underlying logic of this classification scheme is that there are certain skills that are required before students can be expected to read well and these skills are related to being able to read words and understand what they mean.

The outcome measures that the reviewers included within this class are generally subtests within the general reading standardized assessments (examples of standardized assessments with reading subtests include the Gates MacGinitie Reading Test, the Wide Range Achievement Test reading section, the Comprehensive Test of Basic Skills, the Woodcock-Johnson Psycho-Educational Battery, and the Test of Word Reading Efficiency). Examples of decoding outcome measures include Word Identification, Word Attack, Letter Identification, DIBELS fluency assessments, Word Recognition, and Vocabulary tests. Also included in this category were subtests focused on such topics of consonant sounds, short vowels, digraphs and combinations, sight words, and nonword decoding. Some of these measures are not standardized; for example, the Morris et al. 1990 study utilizes a word recognition outcome that is not related to any standardized assessment.

While there are numerous ways that these categories could be divided, the goal was to develop a reasonable number of categories that included similar types of outcome measures. In this category, the focus is on the particular sub-skills necessary for young students to become fluent readers.

Reading Comprehension. In this category, the reviewers included results from comprehension subtests of standardized reading achievement tests. The comprehension subtests used in studies in this review are from the Gates MacGinitie Reading Test, the Wide Range Achievement Test, the Comprehensive Test of Basic Skills, the Stanford Achievement Test, and the Woodcock-Johnson Psycho-Educational Battery.

Reading Oral Fluency. Many of the reading-focused studies employed outcome measures examining the ability of students to quickly and accurately read passages out loud. Such outcomes typically required that students read a passage and rated the students based on the numbers of words correctly read. The outcome measures in this class include the following:

curriculum-based oral fluency, basal passages, observational survey – reading level, analytic reading inventory (fluency), as well as others.

Writing. Only six studies in this review employed outcome measures categorized as writing measures. The outcome measures in this class include the following: spelling, observational survey – writing, observational survey – dictation, the spelling subtest of the Wide Range Achievement Test, the number of words written and spelled correctly in a writing sample, and a curriculum-based spelling measure.

Mathematics Global. Only five studies in this review employed outcome measures categorized as mathematics measures. The measures of math achievement in this class include the following: a researcher-developed multiplication test, the math subtest of the Stanford Achievement Test, and the Orleans-Hanna Algebra Prognosis Test.

Search strategy for Identification of Studies

For this systematic review, titles of studies on volunteer tutoring programs were identified using several methods. First, the reviewers searched the C2-SPECTR, EbscoHost Research Database using the following databases: Academic Search Premier; Primary Search; Professional Development Collection; Middle Search Plus; Psychology and Behavioral Sciences Collection; PsycINFO; Sociological Collection, ERIC (Education Resources Information Center), and Proquest Digital Dissertations.

The initial search included the following key words in various combinations and truncations: "Volunteer or mentor or tutor* or tutorial programs," and "elementary or primary education or middle school students or junior high school students or early intervention," and "control or random or experiment or evaluation or program not peer." In order to find any articles that have yet to be updated in the electronic databases, we conducted online reviews of the table of contents of several major journals that are most relevant to our study, including: *Education Next, Education Policy Analysis Archives* (EPAA), *Educational Evaluation and Policy Analysis* (EEPA), *Reading Research Quarterly* (RRQ), and *Review of Educational Research* (RER) for the years 2003-2005. Only studies conducted in the United States with native English speakers were included in the review. The article search was restricted to studies conducted in the United States only for the sake of policy relevance. That is, interventions employed in educational systems in other countries would not likely fit well within the education policy environment in the United States.

The resulting list of articles was augmented by other research studies referenced in four widely-cited reviews on volunteer tutoring listed above: Elbaum et al., 2000; Wasik, 1998; Shanahan, 1998; and Topping and Hill, 1995. The reviewers also consulted with several sources in order to refine the search process, including an information specialist or librarian, a reading specialist, and the Campbell Collaboration's Information Retrieval Policy Brief (Rothstein, Turner, & Lavenberg, 2004). Studies were retrieved primarily from the University of Arkansas Library System, Interlibrary Loan, University Microfilms, and the databases listed above. All study titles and inclusion decisions were documented and managed using Excel software in order to maintain accuracy and consistency among the reviews. When possible, PDF files of all articles were saved in a central network folder; hard copies of these and print-only articles were also kept on file.

The list of study titles generated in this process was then narrowed through a review of the studies' abstracts by at least two reviewers. Once the abstracts were retrieved and reviewed, both reviewers reviewed the full text of all studies chosen. Studies that passed the full-text initial

review were passed into the full-coding stage. After these studies were fully coded (each coding involved at least two reviewers), the final set of studies that met all inclusion criteria was then analyzed and the results were synthesized. If the two reviewers arrived at different conclusions during the coding process, the coders reconciled whether or not to keep the article. The reconciliation process consisted of a meeting in which each coder explained the rationale for retaining or rejecting the article until agreement was reached. If agreement were not reached, the coders would default to retain the article. In the final stage, the lead reviewer settled any disagreements.

With respect to study design, it is commonly accepted that randomized designs are the strongest designs upon which to base causal inferences. Initial exploration of the available literature uncovered numerous randomized field trials. Choices in the review were based upon the premise that we should use the most reliable evidence available. Had we found only a handful of randomized field trials, we would have then chosen to include high-quality quasi-experimental designs. However, because we identified nearly thirty study cohorts within more than 20 studies or reports based on randomized field trials, we made the decision to exclude quasi-experimental designs from the meta-analysis.

Selection of Trials

As shown in table 1 below, there were 1,437 total studies initially identified based on the search terms in search databases – some of these studies were duplicates from multiple searches. Of the total studies identified, there were 969 unique study abstracts to be reviewed.

Table 1: Original Search Terms and Yields

Search Source	Search Terms	Number of Abstracts
C-2 SPECTR	(All indexed fields) tutor OR (all non-indexed fields) tutor AND 1985-2005 AND volunteer OR mentor OR adult	12
EBSCOhost Research Database (using Academic Search Premier; Primary Search; Professional Development Collection; Middle Search Plus; Psychology and Behavioral Sciences Collection; PsycINFO; Sociological Collection)	(Subject terms) "Volunteer or mentor or tutor* or tutorial programs" and "elementary or primary education or middle school students or junior high school students or early intervention" **AND** (Default fields) control or random or experiment or evaluation or program not peer"	268
ERIC	(SU Descriptors All) "Tutor* not peer not college not faculty" and "primary education or elementary or elementary secondary or junior high or middle school students or early intervention" **AND** "control or random or experiment or program or evaluation"	543
Proquest Digital Dissertations	(Key words) "tutor? or tutorial program?" **AND** (AB) "primary education or elementary or middle school students or early intervention" **AND** (AB) "control or random or experiment or program or evaluation) **AND** (date) ">=1985" **AND** (date) "<=2005"	76
Review of Reviews	Articles published after 1985. Four articles reviewed here are: Elbaum, Vaughn, Hughes, & Moody (2000); Shanahan (1998); Topping & Hill (1995); and Wasik, (1998).	167
Hand Searches	*Dates (2003-2005) - Education Next, Education Policy Analysis Archives* (EPAA), *Educational Evaluation and Policy Analysis* (EEPA), *Reading Research Quarterly* (RRQ), and *Review of Educational Research* (RER)	371
	Total Number of Abstracts Retrieved	1,437
	Total Number of Unique Abstracts Reviewed	969

The list of study titles was generated from electronic database searches, hand searches, and the review of reviews. Abstracts were collected for each title and two coders read each abstract to determine whether the study met the inclusion criteria. The reviewers eliminated abstracts based on the following guide:

- The article reported a meta-analysis

- The study did not employ an experimental design (e.g., case study, narrative, brief analysis/report about 1-2 pages in length, or qualitative study)
- The study was conducted prior to 1985
- The tutees resided outside the United States or were non-English speaking (i.e., ESL, LEP)
- The tutors were peer or professional tutors (i.e., not volunteer adult tutors)
- The students fell outside the parameters of grades K-8
- The focus of the intervention did not have at least one academic outcome
- The intervention duration was less than four weeks, or one month
- The student population was specialized (i.e., deaf, blind)

For a study to be eliminated from the pool, the abstract had to include information clearly indicating that the study met one of the exclusion criteria described above. If not enough information was provided in the abstract, the reviewers could not exclude the study and thus passed it to the next round for further analysis.

A majority of articles during this round were excluded for the following reasons: the studies did not employ an experimental design; the tutors were not volunteers; the tutees were not English speaking or resided outside the United States; or the tutees fell outside the parameters of K-8 grades. The screening stage narrowed the pool of studies a great deal; only 233 of 969 studies passed the initial screening of article abstracts.

The two reviewers then put the remaining 233 articles through a full-text initial review. The initial evaluation of the full text was focused on the introduction and methods section of the article. The purpose of this step was to achieve a fuller understanding of the article without having to fully code every retained article. The reviewers first examined the article to make sure the article reported findings of a study rather than a simple program description. Second, we reviewed the "Methods" section to ensure the students were randomly assigned to a treatment and control group. Finally, the reviewers also eliminated articles based on the general exclusion criteria described above as the article introductions often provided better information than was provided by the abstracts. This stage also resulted in a substantial narrowing of the study pool. Only 56 articles passed this stage and were moved into the full coding stage.

Finally, the final full coding of the 56 retained articles also resulted in the exclusion of more than half of the remaining studies. In the end, only 21 studies were included in our meta-analysis, although we had 28 study "cohorts" from those articles since a few of the articles had multiple cohorts requiring separate coding. For example, the Cobb 2000 article reported results for each of three grades separately and did not provide any pooled data; thus we included the results of this study as three separate Cobb 2000 study cohorts.

A total of 35 articles were excluded during the final full-coding phase. Of the 35 excluded articles, 30 were eliminated due to the following reasons: intervention duration; not a true randomized field trial study design; program intervention implemented outside the U.S.; no relevant academic outcomes; tutoring was not face-to-face; and the use of professional rather than volunteer tutors. The other 5 articles excluded from the meta-analysis were eliminated due to quality concerns with the statistics computed in those studies. The 5 studies in question reported insufficient statistics for us to include in our meta-analysis.

We had initially planned to include quasi-experimental designs, but after identifying 21 studies which employed randomized designs (the strongest designs upon which to base causal inferences), we excluded the quasi-experimental designs from the meta-analysis.

Assessment of Methodological Quality

The quality of each study (and its reporting) was assessed according to several characteristics, including: 1) the transparency of the study; that is, the clarity with which the investigators reported the random assignment procedures; 2) the integrity of the random assignment design and whether investigators addressed violations of the design; 3) the existence of high levels of attrition of either tutees or tutors from the sample initially randomized; and 4) the existence of substantial problems with respect to treatment fidelity. The quality of each study was assessed by all four reviewers during coding review sessions.

Of the 21 studies included in the meta-analysis, none had clear problems with student assignment to treatment and control conditions and none had any evidence of problematic attrition. Questions with fidelity were raised in a few of the included studies and these concerns are noted in Table 4, which summarizes the details of each of the included studies. However, the reviewers did not eliminate any of these studies from consideration.

A final study quality problem was related to reporting of study statistics. Six of the excluded studies failed, at least in part, due to the failure to report adequate descriptive statistics so that the reviewers could compute standardized mean difference effect sizes. Two of these studies in particular (Compton, 1992; Meier & Invernizzi, 2001) did report inferential statistics that would have allowed for the computation of the necessary descriptive statistics; however, the values presented were not consistent. The Compton dissertation had a total sample size of 483, but reported standard deviations that did not match either the reported standard errors or the reported *t* value and that would have yielded an effect size of nearly 2 standard deviations. The Meier and Invernizzi article reported means but no standard deviations, and the *F* values reported had degrees of freedom inconsistent with sample sizes in the study.

Data Management & Extraction

Of the 234 articles initially reviewed, we collected 109 articles from online sources, 66 articles were available in the University of Arkansas library on microfiche, microfilm, or in bound periodicals, 58 articles (including dissertations) were requested by inter-library loan, and 1 dissertation was uncollected when the library determined it had exhausted all possible sources to locate a circulating copy. After collecting an electronic or paper copy of each article, we extracted the information from each article using a designed coding form. The entire review team created the coding form. All study coding, data management, and meta-analysis was done using Excel software.

Data Synthesis

We used Hedges' unbiased estimate *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 post measure) for each outcome measure. When means and standard deviations were not reported, we estimated the effect sizes using the procedures recommended by Wilson and Lipsey (2001). When both pre and post measures were available, we subtracted pretest group differences, which in most cases were minimal because of the requirement of random assignment. If a study reported only adjusted posttest means or posttest

means we computed the treatment-control difference.[2] In either case, we used the pooled standard deviation of the posttest scores as our denominator in computing the effect sizes *d.* To get unbiased estimates of the population effect size, we divided by the approximation of Hedges and Olkin (1985): 1 – 3/(4N – 9).

Some of the studies employed a variety of outcome measures to assess program effectiveness. Because math outcomes are qualitatively different from verbal outcomes, we did not calculate the "effect" of each individual study or the "overall effect" of all available studies. However, we did calculate the overall effect on reading measures. To accomplish this, we computed an overall reading effect size for each of the 25 study cohorts for which some type of reading outcome was assessed. Next, to determine whether an intervention had a greater effect in any one area, we conducted separate meta-analyses of key outcome areas, including standardized overall reading, standardized letters and word skills, standardized reading comprehension, measures of oral reading fluency and of writing, as well as standardized math performance. If a study measured a key outcome in several ways, we averaged the effect sizes of the measures to ensure that each study only contributed one data point to the analysis for each key outcome and that no study was unduly "weighted" in the meta-analysis.

Homogeneity Analysis

The homogeneity analysis test determines if variations in the effect sizes are due to sampling error or other factors. The decision to use a fixed effects model or random effects model is based on the homogeneity analysis. The analyses of the overall effects and of the six key outcomes revealed Q-statistics that were not large enough to allow us to reject the null hypothesis of homogeneity. That is, the variability across effect sizes did not exceed what would be expected based on sampling error (Lipsey & Wilson 2001). Therefore, data syntheses in our study employed fixed effects models.

Sensitivity Analysis

Using the Comprehensive Meta Analysis ® software, we tested the extent to which our main results were sensitive to any one study's inclusion in the meta analysis. The "one study removed" analysis presents the average standardized mean difference of all remaining studies after each study, in turn, is removed from the analysis. In the end, the sensitivity analysis revealed that one study, with a very large sample, had a disproportionate impact on the meta-analytic outcomes. Consequently, all results based on the reading outcomes reported here exclude the Ritter-00 study from the sample (according to the sensitivity analysis, the Rittter-00 study did not have a disproportionate impact on the mathematics global outcome; thus, the Ritter-00 study was retained in the sample for the meta-analysis of the mathematics outcomes).

Subgroup Analysis

Subgroup analyses were conducted to compute differential mean effect sizes based on various program characteristics, including: 1) types of tutors; 2) age of tutees; 3) highly structured vs.

[2] Of the 28 study cohorts analyzed here, 12 were analyzed using posttest scores adjusted for pretest differences (equivalent to gain scores); 3 were analyzed using posttest scores statistically adjusted for pretest differences (using ANCOVA); 12 analyzed based on posttest scores only as no pretest scores were provided; and 1 was analyzed based on a mix of posttest only scores and adjusted posttest scores. The one common denominator was, in fact, the denominator of the effect size calculation. In each case, the denominator of the effect size statistic was the pooled posttest standard deviation.

unstructured programs, and 4) publication source. That is, we examined whether published studies were more likely to show positive program effects. These subgroup analyses were conducted only on the four reading outcomes reviewed here as the writing and math outcome domains only had six and five studies, respectively.

Publication Bias

The publication bias was measured with the "trim and fill" procedure, where the Funnel Plot was visually inspected for differences. Additionally, we examined the possibility of publication bias by using publication type as a moderator variable.

Incomplete Reporting of Study Data

The retained randomized field trials for our study provided enough information for effect sizes to be computed.

PostHoc Subgroup and Moderator Analysis

As part of the meta-analysis, we conducted some subgroup and moderator analysis to determine if differences in effect sizes existed due to characteristics within the studies. The moderator analysis included: 1) type of tutor (parent, college, community); 2) grade 1 or other; 3) level of program structure; and 4) publication type.

Table 2: Description of Program Characteristics for 28 Study Cohorts Included in Meta-Analysis

	Tutors Parents/College/Community	Grade 1 included?	Reading focus?	Structured lessons?	Refereed journal?
Allor-04.1	0 / 1 / 0	1	1	1	1
Allor-04.2	0 / 1 / 0	1	1	1	1
Baker-00	0 / 0 / 1	1	1	0	1
Cobb-01.1	0 / 1 / 0	1	1	0	1
Cobb-01.2	0 / 1 / 0	0	1	0	1
Cobb-01.3	0 / 1 / 0	0	1	0	1
Cook-01.1	0 / 1 / 0	1	1	0	0
Cook-01.2	0 / 1 / 0	0	1	0	0
Cook-01.3	0 / 1 / 0	0	1	0	0
Erion-94	1 / 0 / 0	0	1	0	0
Mahoney-86	1 / 0 / 0	0	0	1	0
Mayfield-00	0 / 1 / 0	1	1	1	0
McKinney-95	0 / 1 / 0	1	0	0	0
Mehran-88	1 / 0 / 0	1	1	1	1
Miller-94	1 / 0 / 0	0	1	0	0
Morris-90.1	0 / 0 / 1	0	1	1	1
Morris-90.2	0 / 0 / 1	0	1	1	1
Nielson-92	0 / 0 / 1	0	1	1	0
Parham-94.1	0 / 0 / 1	0	0	1	0
Parham-94.2	0 / 0 / 1	0	0	0	0
Powell-Smith-00	1 / 0 / 0	0	1	1	1
Pullen-04	0 / 1 / 0	1	1	1	1
Rimm-Kaufman-99	0 / 0 / 1	1	1	1	1
Ritter-00	0 / 1 / 0	0	0	0	0
Vadasy-97a	0 / 0 / 1	1	1	1	1
Vadasy-97b	0 / 0 / 1	1	1	1	1
Vadasy-00	0 / 0 / 1	1	1	1	1
Weiss-89	0 / 0 / 1	1	1	0	0
Total	**5 / 12 / 11**	**14**	**23**	**15**	**15**

4. Results

Description of Included Studies

In the end, the search yielded 21 unique articles, reports, or dissertations; there are 28 unique "study cohorts" or "studies" identified in these 21 reports, as some reports focused on multiple cohorts analyzed separately. Each of the 28 studies is described in some detail in Table 4 in Section 9. The evidence base described here relies upon a sample of 1,676 study participants, 873 of whom were in the tutoring treatment groups and 803 of whom were in the control groups.

Outcome Measures. The volunteer tutoring programs reviewed here employed a variety of outcome measures to assess program effectiveness. After considering the measures in each of the included studies, the reviewers chose to include seven categories of outcome measures in the meta-analysis. The first category is overall reading, for which the evidence is based on the 25 studies (total sample of 1,462 students) which assessed some type of reading measure. However, the sensitivity analysis revealed an outlier study which we removed. Thus, the sample for analysis included 24 studies and a total sample of 1,077 students. After examining overall program effects, we then turn to analyses of the following six specific academic domains (described in detail in the methods section):

(1) Reading global: evidence based on 14 studies with a total sample size of 819.
(2) Reading letters and words: evidence based on 15 studies with a total sample size of 798.
(3) Reading comprehension: evidence based on 8 studies with a total sample size of 546.
(4) Reading oral fluency: evidence based on 12 studies with a total sample size of 635.
(5) Writing: evidence based on 6 studies with a total sample size of 228.
(6) Mathematics global: evidence based on 5 studies with a total sample size of 643.

There is overlap among the studies and samples described above. That is, many of the same studies with reading global outcomes, for example, also assessed outcomes in the reading oral fluency domain.

Types of Tutors. Volunteer tutoring programs generally draw on a variety of sources for tutors. However, there are a few distinctive program types that we separate for analysis here. Some programs train parents as tutors to help their own children. These programs are different from those that train college-age tutors to work with younger students; often these college-age students are in the *America Reads* program or are pre-service teachers. Finally, the remainder of the programs reviewed here used community volunteers from a variety of ages, ranging from older high school students to senior citizens. Programs that used a combination of these tutors are placed in the community volunteer category. In our sample of 28 study "cohorts", 5 were from programs using primarily parents (study sample = 338), 12 were from programs using primarily college-age tutors (study sample = 899), and the remaining 11 were from programs using community volunteers across a variety of ages (study sample = 439).

Age/Grade Level of Tutees. One might conjecture that volunteer tutoring programs work better or worse for older or younger students. Consequently, we divided up our sample of tutoring programs into those that served the youngest students (grade 1) and those that served older students (grade 2 and above). In our sample of 28 study "cohorts", 14 were focused on students in first grade (study sample = 770) and the remaining 14 were focused on older students (study sample = 906).

Program Focus on Reading. Most of the programs included a specific focus on including reading skills; such programs might reasonably be expected to have stronger effects on reading scores than programs without such focus. Consequently, we divided up our sample of tutoring programs into those that focused on reading and those that had a more general academic focus. In our sample of 28 study "cohorts", 23 were focused on reading (study sample = 1,033), while five were not (study sample = 643). We originally planned to use program focus as a moderator variable. However, because only one included study with reading outcomes did not have a reading focus (McKinney-95), we did not conduct subgroup analyses.

Program Structure. Studies were classified as high/low structure depending on the amount of direction and instruction given to the tutors. If the program gave tutors specific lessons and materials to cover, the program was classified as high structure. If the tutees had freedom in selecting the reading materials, but the programs specified how much time in the tutoring session should be spent on each reading activity, the program was also categorized as high structure. Other programs, including some that deliberately were non-directive and provided minimal training to tutors, or programs where tutors and tutees simply read together were classified as low structure.

For example, the Howard Street Tutoring Program, as described by Morris, Shaw and Perney (2000), was classified as structured: "The 3:00-4:00 p.m. tutoring period is carefully planned and work filled, with very few disruptions….a typical 1-hour tutoring lesson takes the following form…" (pp. 136-137). In another case, Vadasy, Jenkins, Antil, and Wayne (1997a) describe an early version of the *Sound Partners* program "The intervention was a set of 100 after-school lessons, each 30 min long…to be used by tutors to teach phonological and early reading skills to first-grade students" (p. 31).

Alternatively, the *Start Making a Reader Today (SMART)* program was decidedly unstructured. "Tutors are provided with a broad framework to use during sessions, rather than specific techniques" (Baker, Gersten, & Keating, 2000, p. 497). They also suggest it is easier to recruit tutors if they are not expected to acquire specialized instructional skills. The paired reading approach and repeated reading approaches, where students select their own materials to read are also classified as non-structured (e.g. Miller, 1998; Weiss, 1989).

In our sample of 28 study "cohorts", 15 were from programs classified as highly structured (study sample = 919), while 13 were not (study sample = 757).

Source of Publication. In the field of systematic reviews, there is a real concern with "publication bias" or "file-drawer bias". These terms refer to the concept that studies showing null effects are less likely to be submitted for publication and less likely to be accepted for publication, all else equal, if submitted. Thus, one might expect that studies published in journals would be more likely to show positive program effects as compared to those disseminated as unpublished reports, conference papers, or student dissertations. Consequently, we distinguished in our sample of studies of tutoring programs that were published in journals as a test of this "bias". In our sample of 28 study "cohorts", 15 were from studies in refereed journals (study sample = 772); the remaining 13 study cohorts were primarily from doctoral dissertations (study sample = 904).

Overall Effect Sizes Across Studies for All Outcomes

We began by examining the overall effects of volunteer tutoring on student reading outcome measures. Twenty-five studies assessed reading measures of one type or another. In eight of these 25 studies, the overall reading score was based on a single reading measure; we computed an average effect size based on multiple reading outcomes for the remaining 17 studies that employed multiple reading outcome measures. Our analysis indicates that volunteer tutoring interventions of the type reviewed here have a significant positive effect on the verbal skills of participating students. Figure 1 shows that the average effect of volunteer tutoring programs on reading outcomes for elementary students is 0.23 using a fixed effects model.

We then conducted sensitivity analyses to examine whether the average effect of volunteer tutoring on reading was disproportionately influenced by the result of any single study. We found that the Ritter-00 study had a sample that comprised approximately 25 percent of the total sample and was more than twice the size of the next largest study sample. The one study removed analysis highlighted in Figure 2 shows that the Ritter-00 did have a substantial negative impact on the average effect size; removing this study increased the average effect size to 0.30. Moreover, the program evaluated in Ritter-00 was unique from the other programs in that there was not a strong academic focus in the program. Given the disproportionate influence of the Ritter-00 study on the average effect size and the unique nature of the program evaluated in Ritter-00, we decided to exclude the study from the domain-specific and subgroup analyses that follow. Figure 3 presents the forest plot and average reading effect size of 0.30 for the set of 24 studies that excludes the Ritter-00 study. Using a 95 percent confidence interval for this effect size, we find that the average effect of volunteer tutoring on reading outcomes ranges from 0.18 to 0.42. The test for heterogeneity produced a Q-value that was not statistically significant ($Q = 17.29$, $p = .80$); thus, all subsequent results are reported using a fixed effects model.

Next, we examined the effect of volunteer tutoring programs on the following specific academic domains (as described in the methods section above): reading global, reading letters and words, reading comprehension, reading oral fluency, writing, and mathematics global.

Reading Global

Thirteen studies assessed outcome measures within the reading global domain; these studies included 195 tutored students in the analysis. As the forest plot for Reading Global indicates (see Figure 4 in Section 10), eight of the studies have positive effect sizes while six have negative effect sizes. However, only two of the most positive effect sizes are statistically different from zero. Overall, the average effect size for this outcome domain is +0.26, an effect that is statistically significant.

Reading Letters and Words

Fifteen studies assessed outcome measures within the letters and words outcome domain; these studies included 403 tutored students in the analysis. As the forest plot for this outcome indicates (see Figure 5 in Section 10), all but two of the studies have positive effect sizes. While only three of these positive effect sizes are statistically different from zero on their own, when all the results are pooled, the effect size is +0.41, an effect that is statistically significant.

Reading Comprehension

Eight studies assessed outcome measures within the reading global domain; these studies included 293 tutored students in the analysis. As the forest plot for Reading Comprehension indicates (see Figure 6 in Section 10), five of the studies have positive effect sizes while only three have negative effect sizes. Only two of these positive effect sizes are statistically different from zero,and the overall effect size is +0.18, an effect that is not statistically significant.

Reading Oral Fluency

Twelve studies assessed outcome measures within the letters and words outcome domain; these studies included 336 tutored students in the analysis. As the forest plot for this outcome indicates (see Figure 7 in Section 10), ten of the twelve studies have positive effect sizes. While only two of these positive effect sizes are statistically different from zero on their own, the pooled effect size for this outcome is +0.30, an effect that is statistically significant.

Writing

Only six studies assessed outcome measures that the reviewers classified within the writing domain; these studies included 111 tutored students in the analysis. As the forest plot for Writing indicates (see Figure 8 in Section 10), all six of these studies have positive effect sizes. While only one of these positive effect sizes is statistically different from zero on its own, the pooled effect size for this outcome is +0.45, an effect that is statistically significant.

Mathematics Global

Only five studies assessed outcome measures that the reviewers classified within the mathematics domain; because these studies were dissertation with large sample sizes, these five studies included a total of 292 tutored students in the analysis. As the forest plot for Mathematics Global indicates (see Figure 9 in Section 10), three of these studies have positive effect sizes and two have negative effect sizes. Only two of the positive effect sizes are statistically different from zero. When these five studies are pooled together, the overall effect size for this outcome domain is +0.27, an effect that is not statistically significant.

Table 2: Mean and Standard Error of Effect Sizes for All Study Outcomes

	N of Studies	N of Tutored Students in Analysis	Effect Size	95% Confidence Interval
Reading – Overall	24	550	0.30*	(.18, .42)
Reading – Global	13	195	0.26*	(.05, .48)
Reading – Letters & Words	15	403	0.41*	(.27, .56)
Reading – Comprehension	8	293	0.18	(-.06, .42)
Reading – Oral Fluency	12	336	0.30*	(.14, .46)
Writing	6	111	0.45*	(.19, .71)
Mathematics	5	292	0.27	(-.18, .72)

* Significantly different from zero, favoring the treatment over the control group, $p < .05$.

Analysis of Impacts for Subgroups of Studies on Reading Outcomes

Next, we examined the possibility of differential effects of different types of volunteer tutoring programs on the reading outcomes. We focus here only on these "subgroup" effects in which there are at least 3 studies in each subgroup. Subgroups examined are described above and include: types of tutors, grade level of tutees, program structure, and publication type.

None of the outcomes had a significant difference in effect size by tutor type. That is, programs using parent, college age, or community tutors did not differ significantly in their effectiveness. Similarly, programs that included Grade 1 were not significantly different from programs for higher grades in their effectiveness. The only significant subgroup difference we found was that highly structured programs had a significant advantage over programs with low structure on the global reading outcome, with an effect size of .59 for structured programs and .14 for unstructured. The other reading outcomes did not differ significantly by amount of program structure. It should be noted that there were only three studies classified as highly structured that used global reading outcomes, and all three studies had the same lead author (Vadasy et al., 1997a, 1997b, 2000).

To assess the possibility of publication bias, we conducted the "trim and fill" procedure, which trims excessively large studies and imputes small studies which may be missing. Figure 10 shows that the 24 studies conform neatly and symmetrically to the shape of the funnel plot, suggesting that there is not a large effect of publication bias on our results. Additionally, the trim and fill procedure was conducted, but there were no excessively large studies to trim. Thus, the observed overall effect size is likely based on an unbiased set of studies.

As an additional test of the possibility of publication bias or "file-drawer bias", we distinguished studies published from dissertations or other unpublished works. The overall trend revealed in this meta-analysis does not indicate publication bias. That is, in each of the reading domains examined here, the effect sizes from the studies published in journals were not significantly larger than those presented in dissertations and unpublished studies.

Table 3: Mean and Standard Error of Effect Sizes, by Publication Type

	Journal Article	Dissertation and Unpublished
Reading Overall (n=24)		
Effect Size	0.36	0.18
Confidence Interval	.21 to .51	-.02 to .39
Number of Studies	14	10
Reading Global (n=13)		
Effect Size	0.43	0.13
Confidence Interval	.05 to .82	-.13 to .39
Number of Studies	6	7
Reading Letters and Words (n=15)		
Effect Size	0.43	0.32
Confidence Interval	.28 to .59	-.07 to .71
Number of Studies	13	2
Reading Comprehension (n=8)		
Effect Size	0.07	0.37
Confidence Interval	-.28 to .42	.06 to .68
Number of Studies	5	3
Reading Oral Fluency (n=12)		
Effect Size	0.31	0.12
Confidence Interval	.15 to .48	-.58 to .81
Number of Studies	10	2

* Significantly different from zero, favoring the treatment over the control group, $p < .05$.

5. Discussion and Conclusions

The objective of this systematic review was to gather, summarize, and integrate the empirical research on the effectiveness of volunteer tutoring programs. The current research base is strong – relatively to that of other educational interventions – in that the review uncovered more than 20 randomized field trials. However, this good news is tempered by the fact that most of these studies employ small samples. Nineteen of the 28 study cohorts in this meta-analysis included 25 or fewer students in the tutoring group; only three of the study cohorts had full study samples (treatment plus control) of more than 100 students. In the end, these 28 study cohorts in these 21 articles reporting on randomized field trials included a full study sample of 1,676 students (873 tutored students and 803 control students). In cases such as this in which there are many studies, most with small samples with very little power to detect program effects, the strength of meta-analysis is that the results of the small studies are pooled and the statistical power is enhanced.

There were numerous academic outcomes assessed in the 21 articles that employed experimental designs to investigate the effectiveness of volunteer tutoring programs. First we analyzed the overall effect of volunteer tutoring programs on all reading outcomes and found a positive and statistically significant positive effect of 0.30 standard deviations. We then grouped the specific outcomes into six domains. Two of these domains were very broad and employed standardized assessments of general skills in reading and math: reading global and math global. The other four outcomes were less broad and focused on specific sub-skills related to reading and language: letters and words, reading comprehension, reading oral fluency, and writing.

The central goal was to examine whether a volunteer tutoring intervention represents a potentially effective strategy for improving academic skills for young students. The answer, according to the existing set of randomized field trials, is a qualified yes. Participation in a volunteer tutoring program results in improved overall reading measures of approximately one-third of a standard deviation. With respect to particular sub-skills, students who work with volunteer tutors are likely to earn higher scores on assessments related to letters and words, oral fluency, and writing, as compared to their peers who were not tutored. The effect sizes connected to these outcome domains were relatively consistent, ranging from 0.26 to 0.45 (See Table 2 above).

A *secondary goal* was to assess whether certain programs are particularly effective. As for the secondary goal, the review reveals that the programs are unique and the individual studies are based on small samples. Furthermore, the programs are small enough so as not to be replicable. Thus, the relevant question may not be "which of these programs are most effective?" Rather, the more important question is "what are the characteristics of effective programs?"

To address this question, the reviewers computed differential mean effect sizes based on various program characteristics, including: 1) types of tutors; 2) age of tutees; and 3) highly structured vs. unstructured programs. These subgroup analyses were conducted only on overall reading and on the four specific reading domains.

However, for the most part, effects were not significantly different across these subgroups. Nonetheless, we can derive some lessons from the characteristics of the programs analyzed. The majority of the studies reviewed here evaluated reading-focused programs delivered to primary age students. Programs did not have to be highly structured to have positive effects, nor did they have to use a particular type of person as a tutor.

In the end, the results of this analysis should serve as one important piece of evidence used by policymakers and educators who are deciding whether to employ volunteer tutoring as a strategy

to improve academic skills for young students. As educators across the country work to meet adequate yearly progress goals in state accountability systems, and as they seek affordable ways to offer additional services to students at risk of not meeting annual academic goals, it would be worthwhile to consider structured, reading-focused volunteer tutoring programs as strategies to improve reading and language skills.

It is also worth noting that the research base for volunteer tutoring, although based mostly on studies with small samples, is useful precisely because there are so many studies that employ experimental designs. This illustrates the power and utility of meta-analysis. While many of the individual studies, standing alone, do not show significant program effects, the overall effect is relatively large and statistically significant in five of the seven outcome domains examined here. As a result, the evidence base in the field benefits from small randomized field trials in which data are reported thoroughly and carefully.

The results of this analysis also reveal that we know very little about the effectiveness of these types of interventions at improving math outcomes. This is disappointing given the important role that early numeracy skills play in later math achievement for students in elementary and middle school. Given this lack of information, it would be useful for educators to develop and implement volunteer tutoring programs focused on early math skills while researchers worked collaboratively to evaluate the effectiveness of these programs. In the end, both practitioners and researchers would be better off as we would begin to learn more about whether this type of intervention might in fact be beneficial for students struggling with early math skills.

6. Plans for Updating the Review

The authors will attempt to update the review approximately every two years. To produce as comprehensive of an update as possible, the researchers plan to conduct more rigorous reviews of the grey literature, contact additional experts in the field of volunteer tutoring, and include studies conducted outside the United States. The researchers also plan to revisit our decision to exclude studies that use quasi-experimental designs.

7. Acknowledgments

The initial investigation into the field of volunteer tutoring was supported under the “Evaluation of the West Philadelphia Tutoring Project” conducted by faculty and staff at the University of Pennsylvania from 1998-2000. The Smith Richardson Foundation provided funding to initiate this review of volunteer tutoring effectiveness. The draft protocol for the pilot review was completed with the help of many colleagues, including participants in the University of Pennsylvania’s Graduate School of Education seminar on conducting systematic reviews and meta-analysis, led by Dr. Rebecca Maynard.

The systematic review has been completed at the University of Arkansas, with the financial support of the Campbell Collaboration. The results of this study do not necessarily represent the official opinion or policy of the Smith Richardson Foundation, the University of Pennsylvania, or the University of Arkansas.

8. Potential Conflict of Interest

Gary Ritter was not involved in the decision whether to include the Ritter 2000 study.

9. Included Studies: Characteristics and References

Table 4: Summary of Characteristics of Included Studies[3]

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
1	**Allor 2004.1 (journal)** *Early literacy tutoring program (Urban schools in the South)*	Gr 1 N = 86 (T=61, C=25)	Education majors (unpaid) or *America Reads* members (stipend)	Three or four 15-20-min sessions per week for 6 mo	Yes	Yes	**Rd-Decoding** WRMT (2), TOWRE (2), DIBELS (1) **Rd-Comprehension** WRMT passage **Rd-Authentic** Oral fluency pre/post gains	Two cohort years considered separately; this is cohort 1
2	**Allor 2004.2 (journal)** *Early literacy tutoring program (Urban schools in the South)*	Gr 1 N =157 (T=76, C=81)	Education majors (unpaid) or *America Reads* members (stipend)	Three or four 15-20-min sessions per week for 6 mo	Yes	Yes	**Rd-Decoding** WRMT (2), TOWRE (2), DIBELS (2) **Rd-Comprehension** WRMT passage **Rd-Authentic** Oral fluency pre/post gains	Two cohort years considered separately; this is cohort 2

[3] The full citations for these 28 studies are presented in a reference list following the table.

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
3	**Baker 2000 (journal)** *SMART – Start Making a Reader Today (Oregon)*	Gr 1 then 2 N = 84 (T=43, C=41)	Community (unpaid)	Two 30-min sessions per wk for two years	Yes	No	**Rd-Decoding** WRMT word ID **Rd-Comprehension** WRMT word, passage (2) **Rd-Authentic** Oral fluency (2) Adjusted posttest	Only the second year findings are included
4	**Cobb 2000.1 (journal)** *Play and phonological awareness activities by preservice teachers.* *(Midwest city)*	Gr 1 N = 18 (T=9, C=9)	Preservice teachers (unpaid)	Two 45-min sessions for 10 wks	Yes	No	***Grade 1*** **Rd-Global** GRAT **Rd-Decoding** GRER (4) posttest *t* values	Three grade levels had different outcomes, and are reported separately; **this is grade 1**. No SDs reported, so effect sizes transformed from reported *t* values to Cohen's *d* [4]

[4] $d = t\ \mathrm{sqrt}(1/n_E + 1/n_C)$ (Glass, McGaw, & Smith, 1981)

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
5	**Cobb 2000.2 (journal)** *Play and phonological awareness activities by preservice teachers.* *(Midwest city)*	Gr 2 N = 20 (T=12, C=8)	Preservice teachers (unpaid)	Two 45-min sessions for 10 wks	Yes	No	***Grades 2 & 3*** **Rd-Global** GRAT **Rd-Decoding** GVOC **Rd-Comprehension** GCOMP posttest *t* values	Three grade levels had different outcomes, and are reported separately; **this is grade 2**. No SDs reported, so effect sizes transformed from reported *t* values to Cohen's *d*
6	**Cobb 2000.3 (journal)** *Play and phonological awareness activities by preservice teachers.* *(Midwest city)*	Gr 3 N = 18 (T=9, C=9)	Preservice teachers (unpaid)	Two 45-min sessions for 10 wks	Yes	No	***Grades 2 & 3*** **Rd-Global** GRAT **Rd-Decoding** GVOC **Rd-Comprehension** GCOMP posttest *t* values	Three grade levels had different outcomes, and are reported separately; **this is grade 3** No SDs reported, so effect sizes transformed from reported *t* values to Cohen's *d*
7	**Cook 2001.1 (dissertation)** *Minimally trained tutors using America Reads materials (suburb of Phoenix, AZ)*	Gr 1 N = 26 (T=12, C=14)	University students (some unpaid, some work study)	Two 45-min sessions per wk for 7 months	Yes	No	**Rd-Global** WRAT3 posttest	Problems with tutor attrition and undependable tutors Reports separated by grade so that data were analyzed as three separate cohorts; **this is grade 1**

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
8	**Cook 2001.2 (dissertation)** *Minimally trained tutors using America Reads materials (suburb of Phoenix, AZ)*	Gr 2 N = 17 (T=7, C=10)	University students (some unpaid, some work study)	Two 45-min sessions per wk for 7 months	Yes	No	**Rd-Global** WRAT3 posttest	Problems with tutor attrition and undependable tutors Reports separated by grade so that data were analyzed as three separate cohorts; **this is grade 2**
9	**Cook 2001.3 (dissertation)** *Minimally trained tutors using America Reads materials (suburb of Phoenix, AZ)*	Gr 3 N = 17 (T=11, C=6)	University students (some unpaid, some work study)	Two 45-min sessions per wk for 7 months	Yes	No	**Rd-Global** WRAT3 posttest	Problems with tutor attrition and undependable tutors Reports separated by grade so that data were analyzed as three separate cohorts; **this is grade 3**
10	**Erion 1994 (dissertation)** *Parent tutoring with flash cards and reading (rural NW Pennsylvania)*	Gr 2 N = 24 (T=12, C=12)	Parents (unpaid)	Five 15-min sessions per wk for 6 wks	Yes	No	**Rd-Authentic** oral reading fluency pre/post gains	
11	**Mahoney 1986 (dissertation)** *Parent tutoring in mathematics (Lakewood OH)*	Gr 3 N = 150 (T=75, C=75)	Parents (unpaid)	Five 30-min. sessions per wk for 4 wks	No	Yes	**Mathematics** 50-item multiplication test; posttest	Random assignment at the classroom level; test was not standardized

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
12	**Mayfield 2000 (dissertation)** *Edmark Reading Program - structured tutoring (Rural northern Louisiana)*	Gr 1 N = 62 (T=31, C=31)	*America Reads* members	Five 15-min sessions per week for 1 semester	Yes	Yes	**Rd-Decoding** WRMT (3) **Rd-Comprehension** WRMT passage pre/post gains [has adjusted post also]	One decoding measure (word attack) reported wrong values for SD-T, so effect size was computed with SD-C rather than SD-pooled. The authentic measure was too closely linked to the curriculum and was not included
13	**McKinney 1995 (dissertation)** *Leap Frog Program: After-school tutorial program at a church (Rural NE Mississippi)*	Gr 1,2 N = 44 (T=20, C=24)	University students (unpaid)	Four 1-hr sessions per wk for 22 wks	No	No	**Rd-Global** Stanford-Reading **Mathematics** Stanford-Mathematics pre/post gains	Outcomes were reported as percentiles
14	**Mehran 1988 (journal)** *Reading Made Easy parent tutoring (small western city)*	Gr 1 N = 76 (T=38, C=38)	Parents (unpaid)	Three 15-min sessions per week for school year	Yes	Yes	**Rd-Global** WJPEB, CTBS **Rd-Decoding** WJPEB (2), CTBS (2), Harrison (4) **Rd-Comprehension** WJPEB, CTBS posttest	Treatment fidelity – most parents reported under 4 sessions a month. Pretest given for CTBS only, but not with the same subscores as posttest, so gains were not computed

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Structured?	Outcome Measures	Concerns/Comments
15	**Miller 1994 (dissertation)** *Paired Reading parent tutoring (Midwest district)*	Gr 2-4 N = 52 (T=26, C=26)	Parents (unpaid)	Four 10-15 min sessions per wk for 10 wks	Yes	No	**Rd-Global** GORT-D, Pre/post gain scores	Treatment fidelity – parents were to tape tutoring sessions, but few did
16	**Morris 1990.1 (journal)** *Howard Street Tutoring Model; after school program (Illinois)*	Gr 2-3 N = 34 (T=17, C=17)	Community Volunteers, various ages (unpaid)	Four 30-min sessions per wk for entire school year	Yes	Yes	**Rd-Decoding** Word recognition measures (3) **Rd-Authentic**. Basal passages **Writing** Spelling correct, qualitative Gain scores only[5]	The 1986-87 and 1987-88 cohorts are separate studies in the meta-analysis. There are some students from each grade in each cohort. Measures may not be standardized
17	**Morris 1990.2 (journal)** *Howard Street Tutoring Model; after school program (Illinois)*	Gr 2-3 N = 26 (T=13, C=13)	Community Volunteers, various ages (unpaid)	Four 30-min sessions per wk for entire school year	Yes	Yes	**Rd-Decoding** Word recognition measures (3) **Rd-Authentic**. Basal passages **Writing** Spelling correct, qualitative Gain scores only	The 1986-87 and 1987-88 cohorts are separate studies in the meta-analysis. There are some students from each grade in each cohort Measures may not be standardized

[5] This study reported only the gain scores and their standard deviations, but no pre/post correlation. Thus, we were unable to derive the posttest pooled standard deviations. To estimate the posttest pooled standard deviation for our effect size computations, we used the gain score standard deviation, which assumes a pre/post correlation of 0.5 and equal pre/post standard deviations.

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
18	**Nielson 1991 (dissertation)** *Parent and adult volunteer tutoring in reading (rural elem. in Delta, Utah)*	Gr 3 N=43 (T=29, C=14)	Parents, adult volunteers (unpaid)	Sessions per wk not stated; program lasted 9 months	Yes	Yes	**Rd-Comprehension** Stanford Achievement Reading Comprehension posttest	Follow-up findings were not included. Parent and volunteer adults were combined into one treatment group for analysis
19	**Parham 1994.1 (dissertation)** *Before-school tutoring in pre-algebra concepts (suburban school) with trained tutors*	Gr 7 N = 32 (T=16, C=16)	Community volunteers (unpaid)	1 60-min session per wk for 5 wks	No	Yes	**Mathematics** OHAPT posttest	We did not include the same-age tutor findings, only adult. No mention of attrition, despite the 7 a.m. start time. A single control group was divided in half between the two Parham treatment groups.
20	**Parham 1994.2 (dissertation)** *Before-school tutoring in pre-algebra concepts (suburban school) with untrained tutors*	Gr 7 N = 32 (T=16, C=16)	Community volunteers (unpaid)	1 60-min session per wk for 5 wks	No	No	**Mathematics** OHAPT posttest	We did not include the same-age tutor findings, only adult. No mention of attrition, despite the 7 a.m. start time. A single control group was divided in half between the two Parham treatment groups.

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
21	**Powell-Smith 2000 (journal)** *Parents using children's literature or basal readers with guided practice and feedback, and monitoring of treatment fidelity.*	Gr 2 N = 36 (T=24, C=12)	Parents (unpaid)	4 20-min sessions per wk for 5 wks	Yes	Yes	**Rd-Authentic** Curriculum Based Measure (CBM), Test of Reading Fluency (TORF) Pre/post gain scores	Two versions of tutoring were combined, one with basal readers and one with children's literature. Some treatment fidelity concerns
22	**Pullen 2004 (journal)** *Repeated reading, coaching in decoding, and reading new books (N. Central Florida)*	Gr 1 N = 49 (T=25, C=24)	University students (unpaid)	40 15-min sessions in 12 wks	Yes	Yes	**Rd-Decoding** Jump Start (3) pre/post gains, WDRB (2) posttest	
23	**Rimm-Kaufman 1999 (journal)** *Comprehensive reading model emphasizing reading for meaning (Cambridge, MA)*	Gr 1 N = 42 (T=21, C=21)	Community volunteers (unpaid)	3 45-min sessions per week for 8 months	Yes	Yes	**Rd-Decoding** Observational Survey (3) **Rd-Writing** Observational Survey (2) **Rd-Authentic** Observational Survey pre/post gains	

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
24	**Ritter 2000 (dissertation)** *West Philadelphia Tutoring Project; University-based partnership (Philadelphia, PA)*	Gr 2-5 N = 385 (T=196, C=189)	College Students (unpaid)	1 hr per week for entire school year	No	No	**Rd-Global** (SAT-9) **Mathematics** (SAT-9) posttest	Subgroup sample sizes for outcomes (after attrition) were not reported but were estimated from the overall T/C sample size ratio
25	**Vadasy 1997a (journal)** *Sound Partners; 100 scripted lessons on phonological awareness, word ID, text reading, writing (large urban district)*	Gr 1 N = 40 (T=20, C=20)	Community volunteers (stipend)	4 30-min sessions per wk for 27 wks	Yes	Yes	**Rd-Global** WRAT-R **Rd-Decoding** WJPEB, Dolch, Yopp-Singer **Rd-Writing** Writing measure, Spelling (2) **Rd-Authentic** Analytic Reading Inventory posttest	Some problem with tutor consistency
26	**Vadasy 1997b (journal)** *Second year of Vadasy 1997a (large urban district)*	Gr 1 N = 40 (T=20, C=20)	Community volunteers (stipend)	4 30-min sessions per wk for 27 wks	Yes	Yes	**Rd-Global** WRAT-R **Rd-Decoding** WJPEB, Dolch, Bryant, Yopp-Singer **Rd-Writing** Writing measure, Spelling (2) **Rd-Authentic** Analytic Reading Inventory adjusted posttest means	Some problem with tutor consistency. Did not include Lesson Word List outcome – it was too specific to the treatment

	Study / Type / Program Description	Tutees	Tutors	Time / Duration	Reading focus?	Highly Struc-tured?	Outcome Measures	Concerns/Comments
27	**Vadasy 2000 (journal)** *Sound Partners; 100 scripted lessons on phonological awareness, word ID, text reading, writing (large urban district)*	Gr 1 N=46 (T=23, C=23)	Community volunteers (stipend)	4 30-min sessions per wk for school yr	Yes	Yes	**Rd-Global** WRAT-R **Rd-Decoding** WJPEB, Dolch, Bryant,Yopp-Singer **Rd-Writing** Writing measure, Spelling (2) **Rd-Authentic** Analytic Reading Inventory (2) adjusted posttest means	Only immediate posttest results are used - 2nd grade follow-up results are not included
28	**Weiss 1988 (unpublished report)** *Paired Reading (suburban district)*	Gr 3-6 N = 20 (T=11, C=9)	Community (unpaid)	4 20- to 30-min sessions per wk for 11 wks	Yes	No	**Rd-Global** BASIS **Rd-Authentic** CBM pre/post gains	

Reference List – Included Studies (21 articles)

Allor, J., & McCathren, R. (2004). The efficacy of an early literacy tutoring program implemented by college students. *Learning Disabilities Research & Practice, 19*(2), 116-129. *(studies 1 and 2)*

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Reading Research Quarterly 35*(4), 494-519. *(study 3)*

Cobb, J. B. (2001). The effects of an early intervention program with preservice teachers as tutors on the reading achievement of primary grade at risk children. *Reading Horizons, 41*(3), 155-173. *(studies 4, 5, and 6)*

Cook, J. A. (2002). Every moment counts: Pairing struggling young readers with minimally trained tutors (Doctoral dissertation, Arizona State University, 2001). *Dissertation Abstracts International, 62*, 8A. *(studies 7, 8, and 9)*

Erion, R. J. (1994). Parent tutoring, reading instruction and curricular assessment (Doctoral dissertation, Indiana University of Pennsylvania, 1994). *Dissertation Abstracts International,54,* 11A. *(study 10)*

Mahoney, E.J. (1986). The effects of parental involvement on achievement of third-grade students (Doctoral dissertation, University of Akron, 1985). *Dissertation Abstracts International, 47*, 2A. *(study 11)*

Mayfield, L. G. (2000). The effects of structured one-on-one tutoring in sight word recognition of first-grade students at-risk for reading failure (Doctoral dissertation, Louisiana Tech University, 2000). *Dissertation Abstracts International, 61,* 2A. *(study 12)*

McKinney, A. D. (1995). The effects of an after school tutorial and enrichment program on the academic achievement and self-concept of below grade level first and second-grad students (Doctoral dissertation, University of Mississippi, 1995). *Dissertation Abstracts International,56,*6A. *(study 13)*

Mehran, M., & White, K. R. (1988). Parent tutoring as a supplement to compensatory education for first-grade children. *RASE: Remedial & Special Education, 9*(3), 35-41. *(study 14)*

Miller, B.V. (1995). An evaluation of the paired reading program using competency-based training. (Doctoral dissertation, University of Wisconsin - Madison, 1994) *Dissertation Abstracts International, 55*, 7A. *(study 15)*

Morris, D., Shaw, B., & Perney, J. (1990). Helping low readers in grades 2 and 3: An after-school volunteer tutoring program. *Elementary School Journal, 91*(2), 133-150. *(studies 16 and 17)*

Nielson, B. B. (1992). Effects of parent and volunteer tutoring on reading achievement of third grade at-risk students (Doctoral Dissertation, Brigham Young University, 1991). *Dissertation Abstracts International, 52*, 10A. *(study 18)*

Parham, J. W. (1994). An analysis of the effects of tutoring on seventh grade students engaged in the mastery of pre-algebra concepts (Doctoral dissertation, University of Maryland College Park, 1993). *Dissertation Abstracts International, 54,* 11A. *(study 19)*

Powell-Smith, K. A., Shinn, M. R., Stoner, G., & Good, R. H. (2000). Parent tutoring in reading using literature and curriculum materials: Impact on student reading achievement. *School Psychology Review, 29*(1), 5-27. *(study 20)*

Pullen, P.C., Lane, H.B., & Monaghan, M.C. (2004). Effects of a volunteer tutoring model on the early literacy development of struggling first grade students. *Reading Research & Instruction,43*(4), 21-40. *(study 21)*

Rimm-Kaufman, S. E., Kagan, J., & Byers, H. (1999). The effectiveness of adult volunteer tutoring on reading among "at risk" first grade children. *Reading Research & Instruction, 38*(2), 143-152. *(study 22)*

Ritter, G. W. (2000) The academic impact of volunteer tutoring in urban public elementary schools: Results of an experimental design evaluation (Doctoral dissertation, University of Pennsylvania, 2000).*Dissertation Abstracts International, 61,* 3A. *(study 23)*

Vadasy P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997a). Community-based early reading intervention for at-risk first graders. *Learning Disabilities Research and Practice, 12*(1), 29-39. *(study 24)*

Vadasy P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997b). The Effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. *Learning Disability Quarterly*, *20*(2), 126-139. *(study 25)*

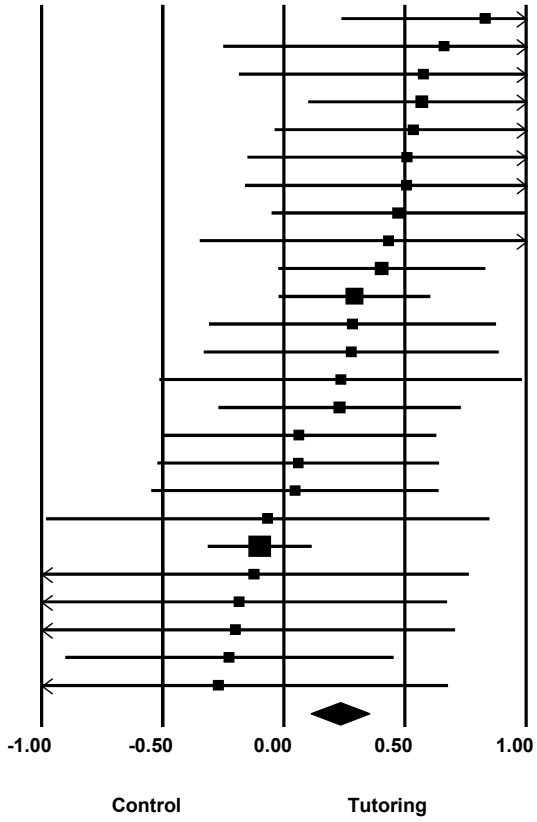
Vadasy, P. F., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at risk for reading disabilities. *Journal of Learning Disabilities, 33*(6), 579-590. *(study 26)*

Weiss, J.A. (1989). *Paired reading with adult volunteer tutors as a reading intervention for students with reading disabilities.* Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989). (ERIC Document Reproduction Services No. ED 305606). *(study 27)*

10. Figures: Forest Plots for All Key Outcomes

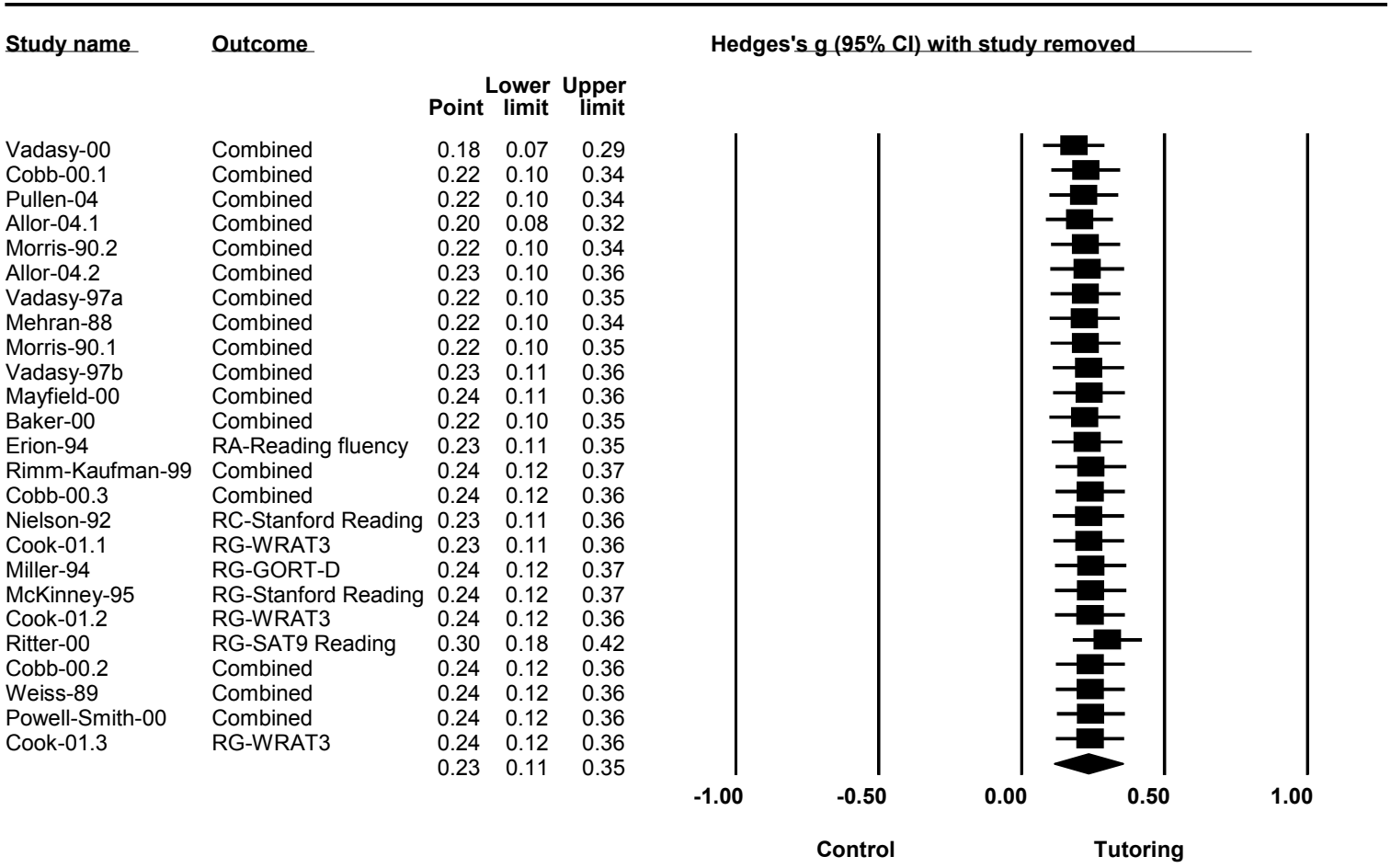
Figure 1: Forest Plot for Effect of Volunteer Tutoring on Children's Reading Outcomes

Study name	Outcome	Hedges's g	Lower limit	Upper limit	Sample size Group-A	Sample size Group-B
Vadasy-00	Combined	0.83	0.24	1.43	23	23
Cobb-00.1	Combined	0.66	-0.25	1.57		
Morris-90.2	Combined	0.58	-0.19	1.34	13	13
Allor-04.1	Combined	0.57	0.10	1.04	61	25
Pullen-04	Combined	0.54	-0.04	1.11	23	24
Vadasy-97a	Combined	0.51	-0.15	1.17	17	18
Morris-90.1	Combined	0.51	-0.16	1.18	17	17
Mehran-88	Combined	0.47	-0.05	1.00	28	28
Erion-94	RA-Reading fluency	0.43	-0.35	1.22	12	12
Baker-00	Combined	0.40	-0.02	0.83	43	41
Allor-04.2	Combined	0.29	-0.02	0.61	76	81
Nielson-92	RC-Stanford Reading	0.28	-0.31	0.88	29	17
Vadasy-97b	Combined	0.28	-0.33	0.89	20	20
Cook-01.1	RG-WRAT3	0.24	-0.51	0.99	12	14
Mayfield-00	Combined	0.23	-0.27	0.73	31	29
Miller-94	RG-GORT-D	0.06	-0.51	0.63	23	23
McKinney-95	RG-Stanford Reading	0.06	-0.52	0.64	20	24
Rimm-Kaufman-99	Combined	0.05	-0.55	0.64	21	21
Cook-01.2	RG-WRAT3	-0.07	-0.98	0.85	7	10
Ritter-00	RG-SAT9 Reading	-0.10	-0.32	0.12	172	156
Cobb-00.3	Combined	-0.12	-1.01	0.77		
Cobb-00.2	Combined	-0.18	-1.04	0.67		
Weiss-89	Combined	-0.20	-1.11	0.71	9	8
Powell-Smith-00	Combined	-0.22	-0.90	0.45	24	12
Cook-01.3	RG-WRAT3	-0.27	-1.22	0.68	11	6
		0.23	0.11	0.35		



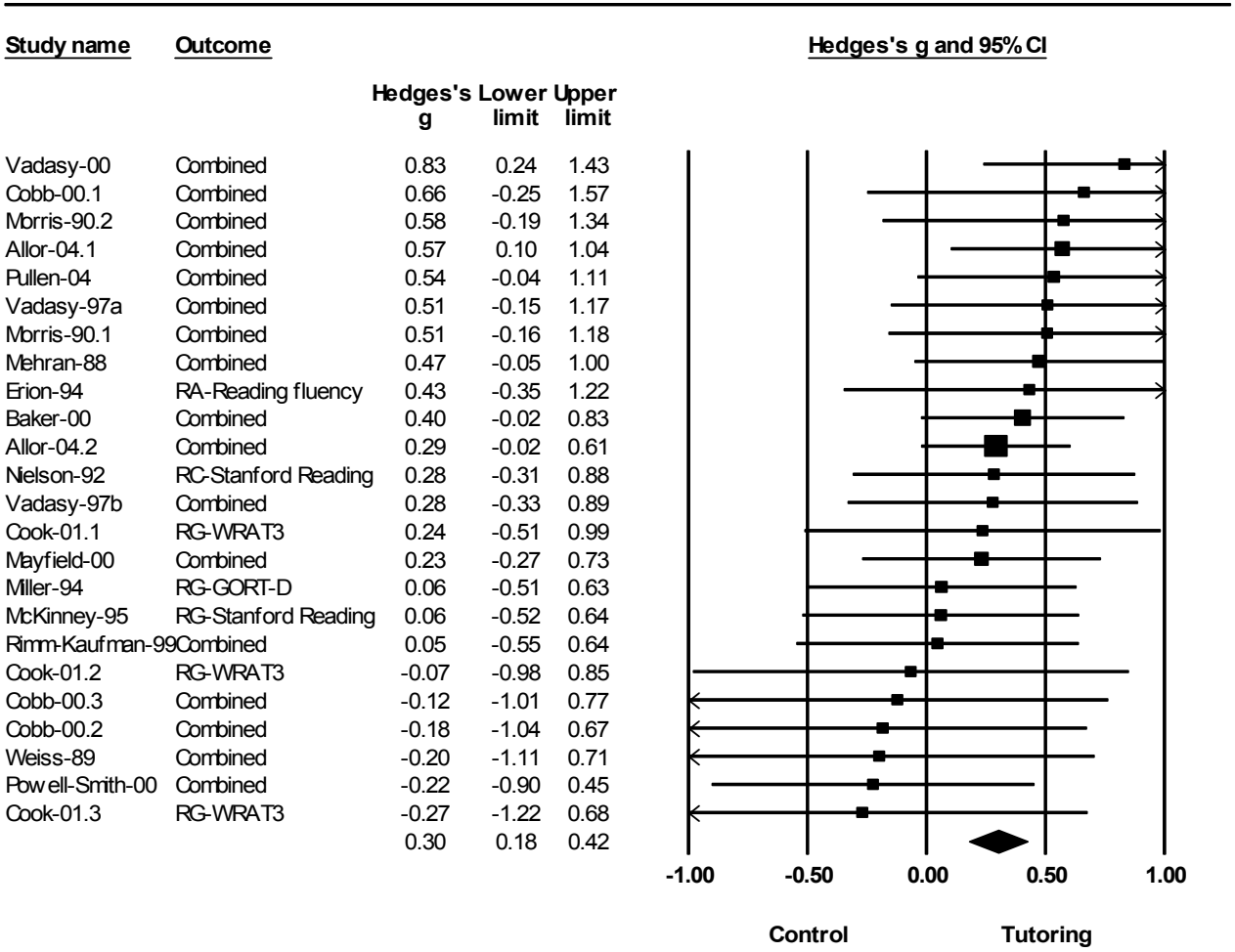
Heterogeneity Statistics for a Fixed Effects Model: Q = 27.20, df = 24, p = .30, and I squared = 11.75

Figure 2: Forest Plot for Effect of Volunteer Tutoring on Children's Reading Outcomes, One Study Removed



Heterogeneity Statistics for a Fixed Effects Model: Q = 27.20, df = 24, p = .30, and I squared = 11.75

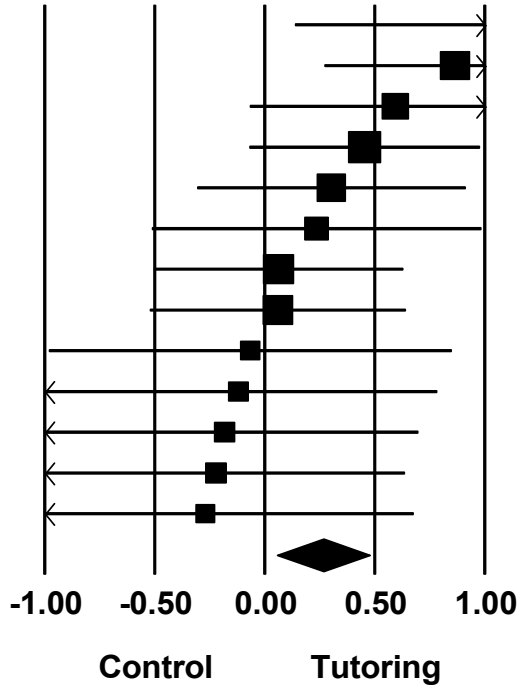
Figure 3: Forest Plot for Effect of Volunteer Tutoring on Children's Reading Outcomes (Ritter-00 excluded)



Heterogeneity Statistics for a Fixed Effects Model: Q = 17.29, df = 23, p = .80, and I squared = 0.00

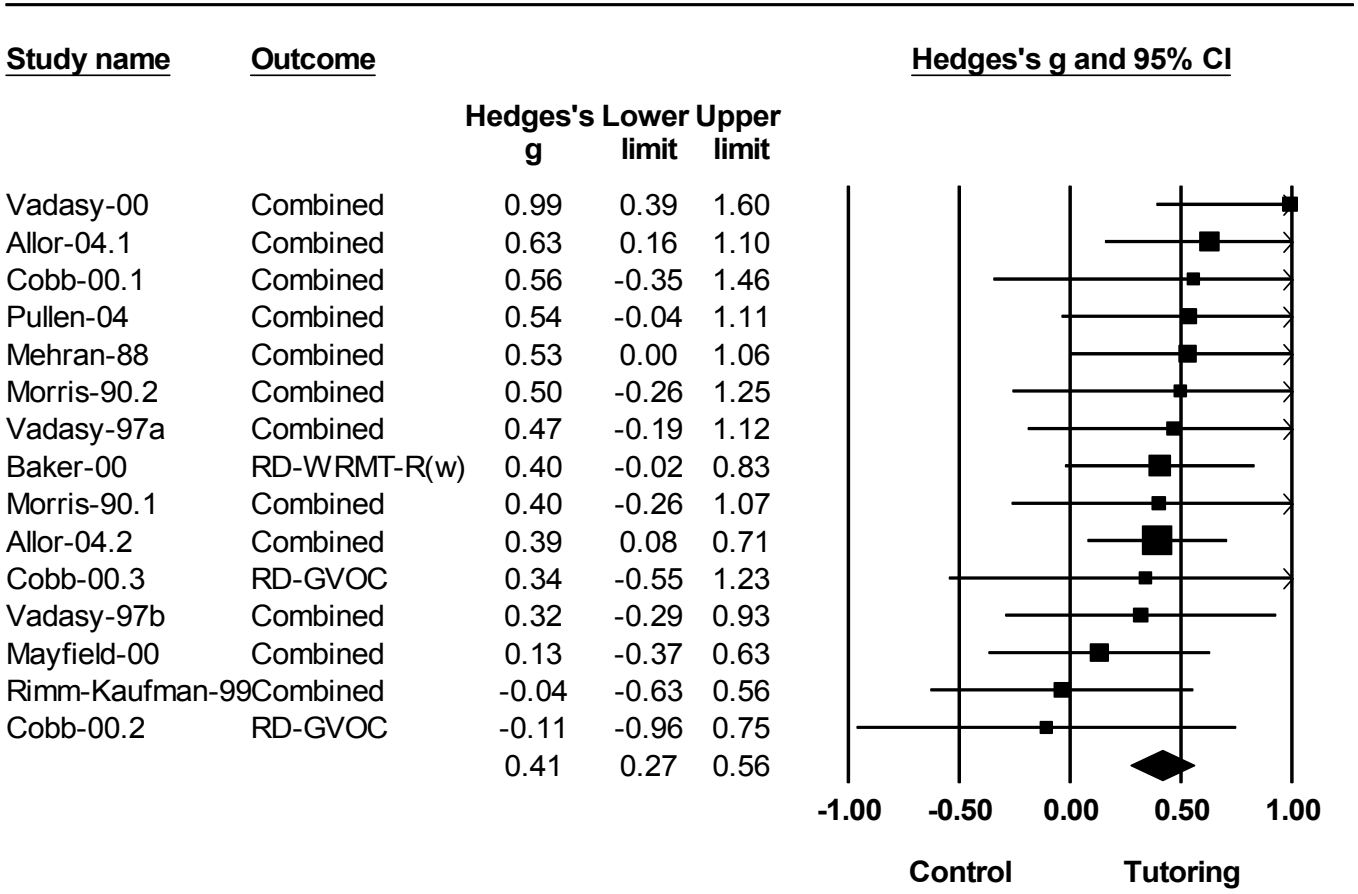
Figure 4: Forest Plot for Reading Global Outcome Measure (Ritter-00 excluded)

Study name	Outcome	Hedges's g	Lower limit	Upper limit
Cobb-00.1	RG-GRAT	1.08	0.14	2.03
Vadasy-00	RG-WRAT-R	0.86	0.27	1.46
Vadasy-97a	RG-WRAT-R	0.59	-0.07	1.26
Mehran-88	Combined	0.45	-0.07	0.98
Vadasy-97b	RG-WRAT-R	0.30	-0.31	0.91
Cook-01.1	RG-WRAT3	0.24	-0.51	0.99
Miller-94	RG-GORT-D	0.06	-0.51	0.63
McKinney-95	RG-Stanford Reading	0.06	-0.52	0.64
Cook-01.2	RG-WRAT3	-0.07	-0.98	0.85
Weiss-89	RG-RGlobal-BASIS	-0.12	-1.02	0.78
Cobb-00.3	RG-GRAT	-0.18	-1.06	0.70
Cobb-00.2	RG-GRAT	-0.22	-1.08	0.64
Cook-01.3	RG-WRAT3	-0.27	-1.22	0.68
		0.26	0.05	0.48



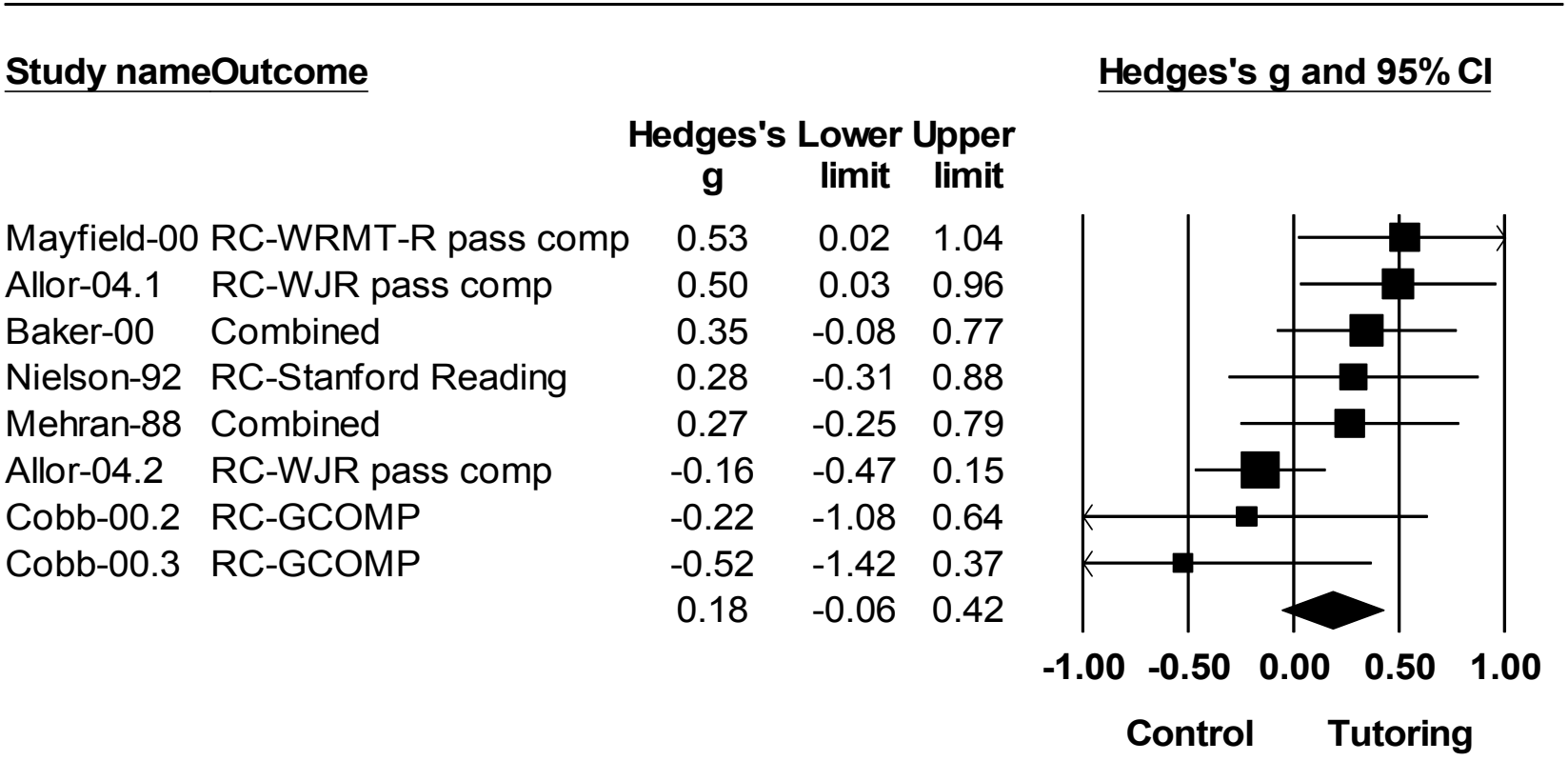
Heterogeneity Statistics for a Fixed Effects Model: Q = 13.84, df = 12, p = .31, and I squared = 13.27

Figure 5: Forest Plot for Reading Letters and Words Outcome Measure



Heterogeneity Statistics for a Fixed Effects Model: Q = 9.83, df = 14, p = .78, and I squared = 0.00

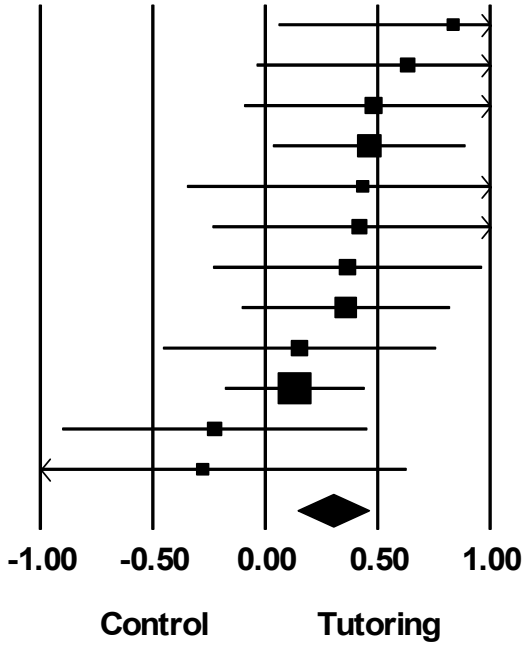
Figure 6: Forest Plot for Reading Comprehension Outcome Measure



Heterogeneity Statistics for a Fixed Effects Model: Q = 12.03, df = 7, p = .10, and I squared = 41.82

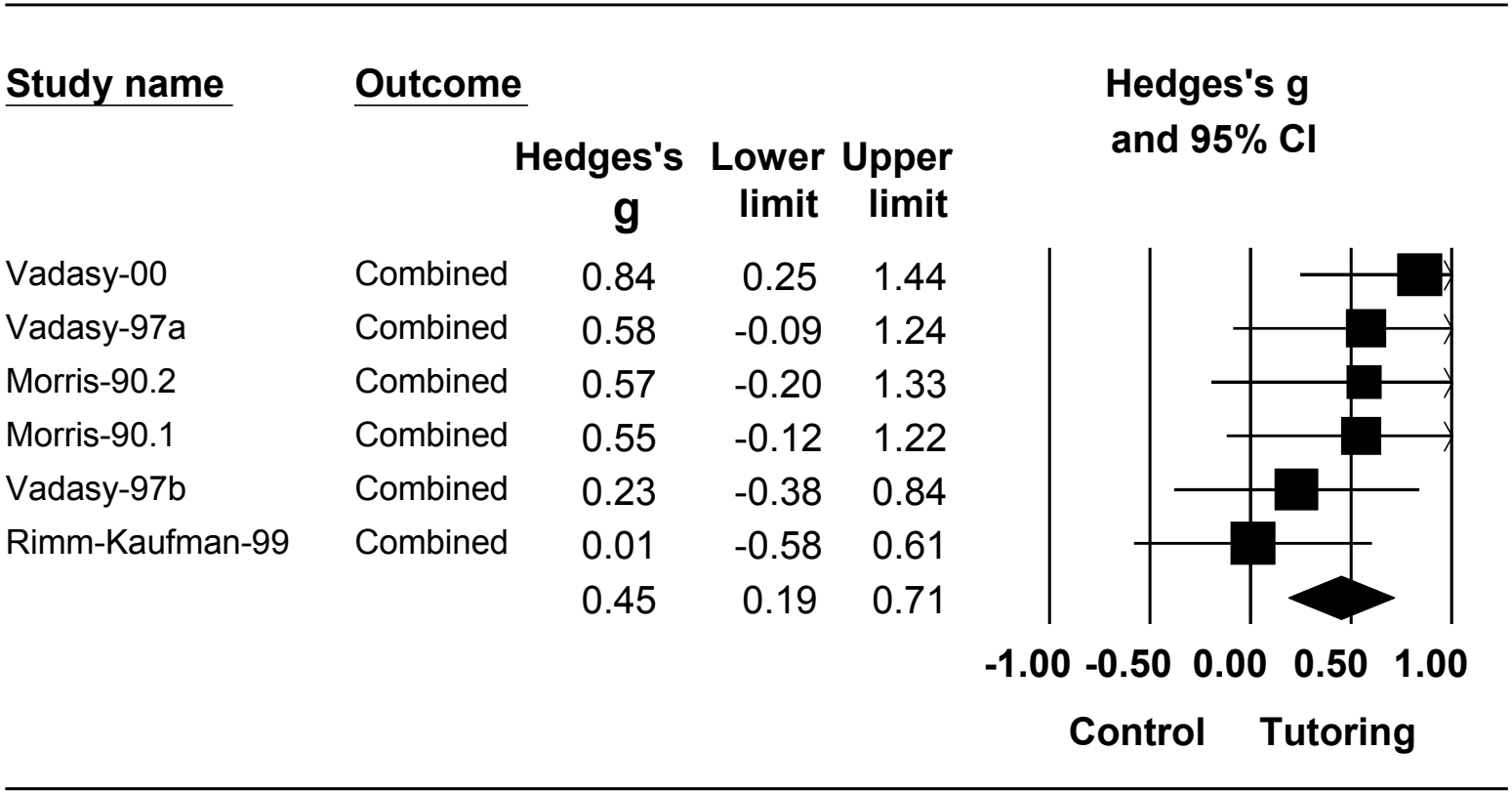
Figure 7: Forest Plot for Reading Oral Fluency Outcome Measure

Study name	Outcome	Hedges's g	Lower limit	Upper limit
Morris-90.2	RA-basal passages	0.84	0.06	1.61
Morris-90.1	RA-basal passages	0.63	-0.04	1.31
Vadasy-00	Combined	0.48	-0.09	1.06
Baker-00	Combined	0.46	0.03	0.89
Erion-94	RA-Reading fluency	0.43	-0.35	1.22
Vadasy-97a	RA-Reading Fluency ARI	0.42	-0.24	1.07
Rimm-Kaufman-99	RA-RdgLevel	0.37	-0.23	0.96
Allor-04.1	RA-Reading fluency	0.36	-0.11	0.82
Vadasy-97b	RA-Reading Fluency ARI	0.15	-0.46	0.76
Allor-04.2	RA-Reading fluency	0.13	-0.18	0.44
Powell-Smith-00	Combined	-0.22	-0.90	0.45
Weiss-89	RA-RFluency-CBM	-0.28	-1.19	0.63
		0.30	0.14	0.46



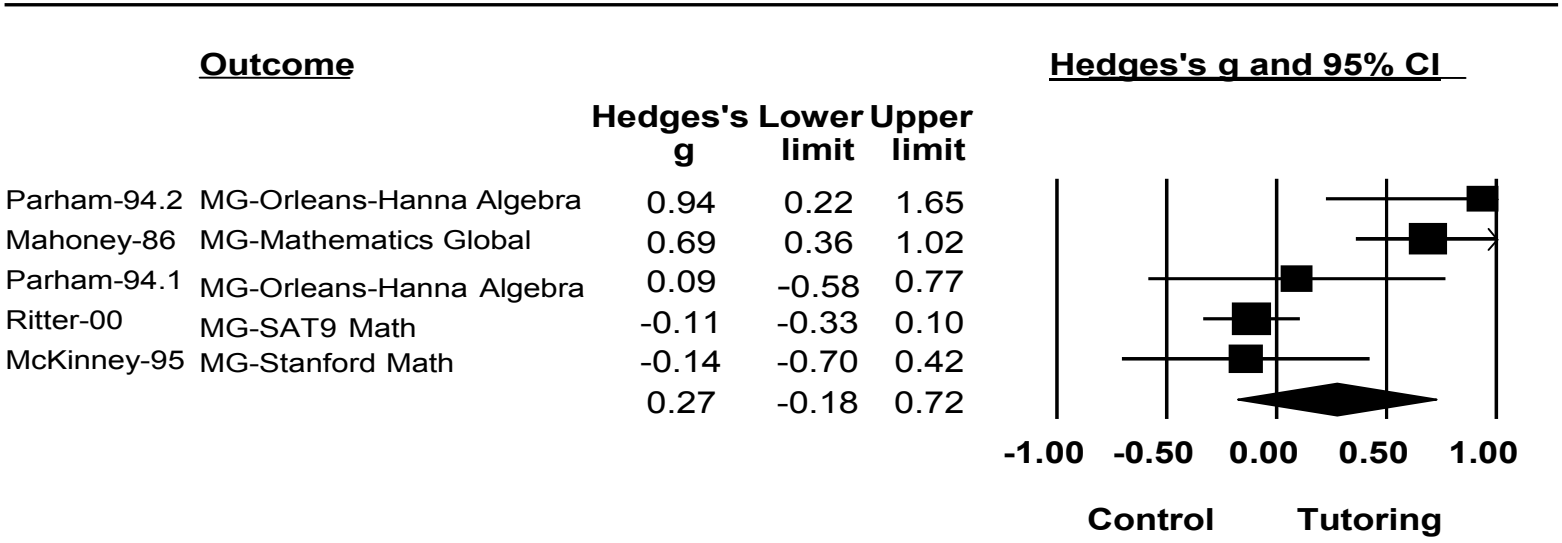
Heterogeneity Statistics for a Fixed Effects Model: Q = 9.24, df = 11, p = .60, and I squared = 0.00

Figure 8: Forest Plot for Writing Outcome Measure



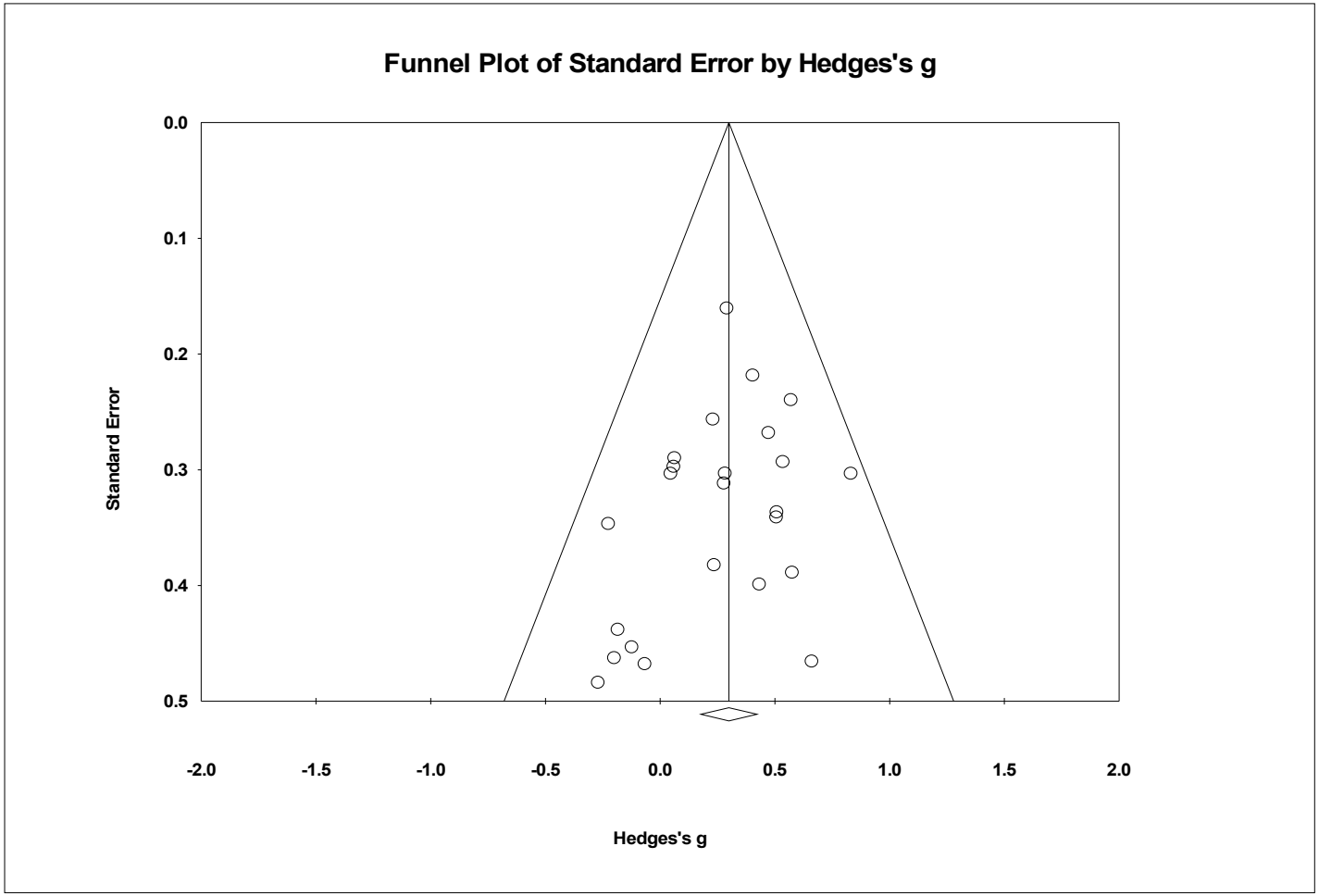
Heterogeneity Statistics for a Fixed Effects Model: Q = 4.57, df = 5, p = .47, and I squared = 0.00

Figure 9: Forest Plot for Mathematics Global Outcome Measure



Heterogeneity Statistics for a Fixed Effects Model: Q = 21.68, df = 4, p < .01, and I squared = 81.55

Figure 10: Funnel Plot of Standard Error by Hedges's g for Overall Reading (24 studies, Ritter-00 excluded)



11. Additional Tables

Table 5: Level of Structure of Various Programs

Study	Coded	Evidence of high/low structure
Allor-04.1,2	High	Tutoring sessions had a game, word-study activities, and book-reading activities
Baker-00	Low	Tutors are provided with a broad framework rather than specific techniques
Cobb-01.1,2,3	Low	Small toys, games, and children's books made available, tutors created activities
Cook-01.1,2,3	Low	Tutors were given the *America Reads* guidebook, but were not required to use any of the strategies.
Erion-94	Low	Parents used repeated reading going consecutively through a basal reader
Mahoney-86	High	Parents given multiplication lessons and worksheets to work with their children
Mayfield-00	High	15 minutes of the *Edmark Reading Program*, a sequenced, repetitive, sight-word approach
McKinney-95	Low	Classroom teachers send work with student to *Leap Frog* after-school tutoring program
Mehran-88	High	Parents administer *Reading Made Easy* lessons
Miller-88	Low	Paired reading- parents read with children alone and simultaneously with books the child selected
Morris-90.1,2	High	60-minute tutoring sessions are carefully planned and work filled with reading, word study, writing
Nielson-92	High	Tutors given stopwatch, flash cards, stories, and logs and trained in how to use them in the tutoring sessions.
Parham-94.1	High	Tutors were trained and provided with instructional materials
Parham-94.2	Low	Tutoring sessions supplemented classroom lessons
Powell-Smith-00	High	Sessions were 20 minutes with 3 components: preview, child reads aloud, choice activities
Pullen-04	High	Sessions were 15 minutes with Gaining Fluency, Measuring Progress, and Reading a New Book,
Rimm-Kaufman-99	High	Each session followed a schedule: familiar material, more challenging, then closing with familiar work.
Ritter-00	Low	The subject matter covered during the session varied by school site and by individual tutor-tutee pairing
Vadasy-97a, 97b, 00	High	*Sound Partners* program: 100 after school lessons, each 30 minutes long
Weiss-89	Low	Paired reading - students selected materials to read with the tutor

Table 6: Components of Effect Size Outcomes by Outcome Domain by Study

Reading: Global

Study	*Measure*	*Unbiased d*
Cobb-01.1	Gates-MacGinitie Reading Test (GRAT)	1.08
Cobb-01.2	GRAT	-0.22
Cobb-01.3	GRAT	-0.18
Cook-01.1	Wide-Range Achievement Test (WRAT)	0.24
Cook-01.2	WRAT	-0.06
Cook-01.3	WRAT	-0.27
McKinney-95	Stanford Achievement Tests (8th ed.)	0.06
Mehran-88	Woodcock-Johnson Psycho-Educational Battery (WJPEB) Comprehensive Test of Basic Skills (CTBS) *Combined mean*	0.66 0.25 0.45
Miller-88	Gray Oral Reading Test-Diagnostic (GORT-D)	0.06
Ritter-00	Stanford Achievement Test (9th ed.)	-0.10
Vadasy-97a	WRAT – Reading	0.59
Vadasy-97b	WRAT – Reading	0.30
Vadasy-00	WRAT – Reading	0.86
Weiss-89	Basic Achievement Skills Individual Screener (BASIS) – Raw score	-0.18

Reading: Letters and Words

Study	*Measure*	*Unbiased d*
Allor-04.2	Woodcock Johnson-Revised (WJ-R) Word ID WJ-R Word Attack Test of Word Reading Efficiency (TOWRE) Real Word TOWRE Nonword Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Phoneme-segmentation fluency *Combined Mean*	0.57 0.85 0.30 0.80 0.62 0.63
Allor-04.2	WJ-R Word ID WJ-R Word Attack TOWRE Real Word TOWRE Nonword DIBELS Phoneme-segmentation fluency DIBELS Nonsense-word fluency *Combined Mean*	0.20 0.66 0.13 0.55 0.44 0.40 0.39
Baker-00	Woodcock Reading Mastery Test-Revised (WRMT-R) Word ID	0.40
Cobb-01.1	GRAT-Initial consonants & clusters (GRER1) GRAT-Final consonants & clusters (GRER2) GRAT-Vowels (GRER3) GRAT-Use of sentence context (GRER4) *Combined Mean*	0.32 0.71 0.95 0.25 0.56
Cobb-01.2	GRAT-Vocabulary (GVOC)	-0.11
Cobb-01.3	GRAT-Vocabulary (GVOC)	0.33
Mayfield-00	WRMT-R Word ID WRMT-R Letter ID WRMT-R Word Attack *Combined Mean*	-0.01 0.26 0.13 0.13

Mehran-88	WJPEB Letter/Word ID	0.56
	WJPEB Word Attack	0.71
	CTBS Word Analysis	0.15
	CTBS Vocabulary	0.33
	Harrison Criterion Referenced Test (HCRT) – Producing Sounds	0.75
	HCRT Consonant Sound	0.62
	HCRT Short Vowels	0.33
	HCRT Digraphs and Combinations	0.77
	Combined Mean	0.53
Morris-90.1	Word recognition (untimed)	0.24
	Basal word recognition	0.68
	Combined Mean	0.46
Morris-90.2	Word recognition (timed/flash)	0.61
	Word recognition (untimed)	0.43
	Basal word recognition	0.75
	Combined Mean	0.60
Pullen-04	Jump Start – Phonological Awareness	0.39
	Jump Start – Sight Words	0.24
	Jump Start – Nonword Decoding	0.82
	WDRB – Letter Word ID	0.38
	WDRB – Word Attack	0.86
	Combined Mean	0.63
Rimm-Kaufman-99	Observational Survey – Letters	0.09
	Observational Survey – Words	0.02
	Observational Survey – Concepts about Print	-0.21
	Combined Mean	0.05
Vadasy-97a	WJ-R – Word Attack	0.35
	Dolch Word Recognition	0.31
	Yopp-Singer Word Segmentation	0.74
	Combined Mean	0.47
Vadasy-97b	WJ-R – Word Attack	0.21
	Dolch Word Recognition	0.15
	Yopp-Singer Word Segmentation	0.41
	Bryant pseudoword	0.54
	Pseudoword list	0.28
	Combined Mean	0.32
Vadasy-00	WJ-R – Word Attack	1.21
	Dolch Word Recognition	0.81
	Yopp-Singer Word Segmentation	0.90
	Bryant pseudoword	1.04
	Combined Mean	0.99

Reading: Comprehension

Study	*Measure*	*Unbiased d*
Allor-04.1	WJ-R Passage Comprehension	0.50
Allor-04.2	WJ-R Passage Comprehension	0.45
Baker-00	WRMT-R Word Comprehension	0.41
	WRMT-R Passage Comprehension	0.28
	Combined Mean	0.34
Cobb-01.2	GRAT-Reading Comprehension (GCOMP)	-0.22
Cobb-01.3	GRAT-Reading Comprehension (GCOMP)	-0.52
Mayfield-00	WRMT-R Passage Comprehension	0.53

Mehran-88	WJPEB Passage Comprehension CTBS Comprehension *Combined Mean*	0.44 0.09 0.27
Nielson-92	Stanford Achievement Test - Comprehension	0.28

Reading: Oral Fluency

Study	*Measure*	*Unbiased d*
Allor-04.1	Curriculum-based oral reading fluency	0.36
Allor-04.2	Curriculum-based oral reading fluency	0.30
Baker-00	Oral reading fluency – 1st grade passage Oral reading fluency – 2nd grade passage *Combined Mean*	0.45 0.47 0.46
Erion-94	Oral reading fluency	0.27
Morris-90.1	Basal passages	0.77
Morris-90.2	Basal passages	1.00
Rimm-Kaufman-99	Observational Survey – Reading Level	0.37
Vadasy-97a	Analytic Reading Inventory (fluency)	0.42
Vadasy-97b	Analytic Reading Inventory	0.15
Vadasy-00	Analytic Reading Inventory – Primary level Analytic Reading Inventory – 1st grade level *Combined Mean*	0.56 0.40 0.48
Weiss-89	Curriculum-based Measurement (CBM) Raw	-0.15

Writing

Study	*Measure*	*Unbiased d*
Morris-90.1	Spelling (correct score) Spelling (qualitative score) *Combined Mean*	0.61 0.83 0.72
Morris-90.2	Spelling (correct score) Spelling (qualitative score) *Combined Mean*	0.83 0.72 0.77
Rimm-Kaufman-99	Observational Survey – Writing Observational Survey – Dictation *Combined Mean*	-0.01 0.18 0.08
Vadasy-97a	WRAT-R Spelling (Standard) Writing sample - Words written Writing sample - Words correctly spelled *Combined Mean*	0.78 0.44 0.37 0.53
Vadasy-97b	WRAT-R Spelling (Raw) Writing sample - Words written Writing sample - Words correctly spelled *Combined Mean*	0.20 0.26 0.18 0.21
Vadasy-00	WRAT-R Spelling Subtest (standard) Curriculum-based spelling measure Words correct (%) on writing measure *Combined Mean*	0.80 0.91 0.81 0.84

Mathematics

Study	*Measure*	*Unbiased d*
Mahoney-86	Multiplication test	0.69
McKinney-95	Stanford Achievement Test (8th ed.)	-0.14
Parham-94.1	Orleans-Hanna Algebra Prognosis Test	.09
Parham-94.2	Orleans-Hanna Algebra Prognosis Test	.96
Ritter-00	Stanford Achievement Test (9th ed.)	-0.12

12. Excluded Studies: Characteristics and References

Table 7: Summary of Characteristics of Excluded Studies

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
1	Abbott, S. P., & Berninger, V. W. (1999). It's never too late to remediate: Teaching word recognition to students with reading disabilities in grades 4-7. *Annals of Dyslexia, 49,* 223-50.	Gr 4-7	School Psychologists and Grad Students in Psych	▪ Tutors were not volunteers ▪ Two types of treatment were compared.
2	Al-Hazza, T. C. (2003) An examination of the effects of the America Reads tutoring program and tutor training on the attitude and academic achievement of urban at-risk minority students (Doctoral dissertation, Old Dominion University, 2002). *Dissertation Abstracts International, 63,* 10A.	Gr K-3	America Reads College Students	▪ Not a Randomized Field Trial (2 treatment and 2 comparison schools matched on SES and achievement)
3	Allen, A., & Chavkin, N. F. (2004). New evidence that tutoring with community volunteers can help middle school students improve their academic achievement. *School Community Journal, 14*(2), 7-18.	Middle school	America Reads College Students	▪ Not a Randomized Field Trial (compared those with more tutoring to those with less)
4	Berninger, V. W., Abbott, R. D., & Whitaker, D. (1995). Integrating low-and high-level skills in instructional protocols for writing disabilities. *Learning Disability Quarterly, 18*(4), 293-309.	Elem Students		▪ Tutors were not volunteers (two experienced clinicians).

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
5	Bradley, K. L. (2002). The effects of the help one student to succeed (HOSTS) program on the reading achievement of at-risk 4th and 5th grade elementary students (Doctoral dissertation, Old Dominion University, 2001). *Dissertation Abstracts International, 63,* 2A.	Gr 4-5	Community Volunteers (various ages)	▪ Not a Randomized Field Trial (39 students in HOSTS matched on SES to comparison group)
6	Burns, M. K., Senesac, B. V., & Symington, T. (2004). The effectiveness of the HOSTS program in improving the reading achievement of children at-risk for reading failure. *Reading Research & Instruction, 43*(2), 87-104.	Gr K-5	Community Volunteers (various ages)	▪ Not a Randomized Field Trial (compared 6 schools already using program to 4 matched schools not using program)
7	Cobb, J. B. (2000). Impact of major field of study on tutors' performance: A literacy intervention program for at risk fourth graders. *Journal of Reading Education, 25*(3), 22-31.	Gr 4	Prospective Teacher and College Athlete	▪ No control group; comparison between two treatments with two different types of tutors
8	Community Service Society of New York. (2000). *The experience corps: An intergenerational program flourishes in the South Bronx*. New York: Urban Agenda Issue Brief. (ERIC Document Reproduction Services No. ED476328)	Gr 1	Retirees	▪ Inadequate reporting of statistics; provided only means (appears to refer to article by Meier and Invernizzi, 2001)
9	Compton, G. L. (1992) The Reading Connection: A leadership initiative designed to change the delivery of educational services to at-risk children (Doctoral dissertation, Western Michigan University, 1992). *Dissertation Abstracts International, 53,* 4A.	Gr 1	Teacher Education Students	▪ Inadequate reporting of statistics; only table reporting descriptive statistics has numbers that are not feasible (degrees of freedom)

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
10	Rabiner, D. L., Malone, P. S., and the Conduct Problem Prevention Research Group (2004). The impact of tutoring on early reading achievement for children with and without attention problems. *Journal of Abnormal Child Psychology, 32*(3), 273-284.	Gr 1	Paraprofessionals	▪ Tutors were not volunteers; paraprofessionals who received 40 hours of training in the Wallach and Wallach method
11	Edwards, H. M. (2001). The effects of tutorial and mentoring initiatives employed by military/school partnerships on selected improvement variables for at-risk elementary students in Bexar County, Texas (Doctoral dissertation, Texas A&M University, 2000). *Dissertation Abstracts International, 61,* 11A.	Gr 3-5	Mentors from a military installation in Texas, ages 20-60	▪ Not a Randomized Field Trial; randomly selected a subset of the tutored group for analysis, compared to a random subset from non-tutored population
12	Fantuzzo, J. W., Davis, G. Y., & Ginsburg, M. D. (1995). Effects of parent involvement in isolation or in combination with peer tutoring on student self-concept and mathematics achievement. *Journal of Educational Psychology, 87*(2), 272-281.	Gr 4-5	Peer Tutoring	▪ Not a volunteer tutoring program; instead, a peer tutoring program
13	Ferrier, M. H. (1994). Mentoring: Its effects on at-risk elementary students, mentors and teachers of mentored students (Doctoral dissertation, Texas A&M University, 1993). *Dissertation Abstracts International, 54*, 8A.	Gr 3-5	Mentors from a Kelley Air Force Base in Texas	▪ Not a Randomized Field Trial; randomly selected a subset of the tutored group for analysis, compared to a random subset from non-tutored population

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
14	Fowler, M. C., Lindemann, L. M., Thacker-Gwaltney, S., & Invernizzi, M. (2002). *A second year of one-on-one tutoring: An intervention for second graders with reading difficulties. CIERA Report.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement. (ERIC Document Reproduction Service No. ED468418).	Gr 2	Community Volunteers and Federal Work Study Graduate Students	▪ Not a Randomized Field Trial; second grade teachers and principals influenced the assignment of students to intervention (N=26 children assigned to treatment were perceived as needier despite similar pretest scores)
15	Invernizzi, M., Rosemary, C., & Juel, C., & Richards, H. C. (1997). At-risk readers and community volunteers: A 3-year perspective. *Scientific Studies of Reading, 1*(3), 277-300.	Gr 1	Community Volunteers (various ages)	▪ Not a Randomized Field Trial (compared those with more tutoring to those with less)
16	Jason, L. A., Danner, K. E., & Kurasaki, K. S. (1993). A 1-year follow-up of a preventive program for high-risk transfer children. *Journal of Emotional & Behavioral Disorders, 1*(4), 215-221.	Gr 3-5 (trans-fers)	University undergraduates and parents	▪ No relevant outcome measure; only outcome is reading grades
17	Jason, L. A., Johnson, J. H., & Danner, K. E. (1993). A comprehensive, preventive, parent-based intervention for high-risk transfer students. *Prevention in Human Services, 10*(2), 27-37.	Gr 3-5 (trans-fers)	University undergraduates and parents	▪ No relevant outcome measure; only outcome measure is coping strategy
18	Jason, L. A., Weine, A. M.,& Johnson, J. H. (1993). The school transitions project: A comprehensive preventive intervention. *Journal of Emotional & Behavioral Disorders, 1*(1), 65-70.	Gr 3-5 (trans-fers)	University undergraduates and parents	▪ Inadequate reporting of statistics; descriptive statistics not reported in online article; need information on reading global measure
19	Jason, L. A., Betts, D., &; Johnson, J. (1989). An evaluation of an orientation plus tutoring school-based prevention program. *Professional School Psychology, 4*(4), 273-284.	Gr 3-5 (trans-fers)	University undergraduates and parents	▪ Inadequate reporting of statistics; descriptive statistics not reported

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
20	Johnson, J. H., Jr. (1992). An empirical study of elementary transfer student adjustment: Developing a model of transfer student adjustment using path analysis and classifying transfer student subtypes based on cluster analysis (Doctoral dissertation, DePaul University, 1991). *Dissertation Abstracts International, 52,* 10B.	Gr 3-5 (trans-fers)	University undergraduates and parents	▪ No relevant outcome measure; outcomes include school grades in reading and math, as well as several behavioral measures
21	Juel, C. (1996). What makes literacy tutoring effective? *Reading Research Quarterly, Vol 31*(3), 268-289.	Gr 1-2	College Student-Athletes	▪ Not a Randomized Field Trial; QED with non-equivalent comparison group (tutored low-achievers, compared to the rest of students)
22	Lauren, L., & Allen, L. (1999). Factors that predict success in an early literacy intervention project. *Reading Research Quarterly, 34*(4), 404-424.	Gr 1-4	Preservice teachers	▪ Not a Randomized Field Trial; QED with constructed control group (matched on pretest scores)
23	Law, M., & Kratochwill, T. R. (1993). Paired reading: An evaluation of a parent tutorial program. *School Psychology International, 14*(2), 119-147.	Gr 2-4	Parents trained in paired reading	▪ Not a Randomized Field Trial; pre-post comparison with no control group
24	Leach, D. J., & Siddall, S. W. (1990). Parental involvement in the teaching of reading: A comparison of hearing reading, paired reading, pause, prompt, praise, and direct instruction methods. *British Journal of Educational Psychology, 60*, 349-355.	Gr 1	Parents	▪ Study not conducted in United States

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
25	Meier, J. D., & Invernizzi, M. (2001). Book buddies in the Bronx: Testing a model for America Reads. *Journal of Education for Students Placed at Risk, 6*(4), 319-333.	Gr 1	Retirees	▪ Inadequate reporting of statistics; provided only means (and F-statistics that were not feasible based on reported degrees of freedom).
26	Meyer, B. J., Middlemiss, W., & Theodorou, E. (2002). Effects of structure strategy instruction delivered to fifth-grade children using the Internet with and without the aid of older adult tutors. *Journal of Educational Psychology, 94*(3), 486-519.	Gr 5	Retirees	▪ Not a study of a face-to-face tutoring program (program is internet based)
27	Rembert, W. I. (1989). Cross-age tutoring and young children's spatial problem-solving skills in a Logo programming environment. Ph.D. dissertation, The University of North Carolina at Greensboro		Various ages, including college students	▪ Tutoring program did not meet minimum duration of one month (program lasted only 3 weeks)
28	Rogers, S. F. (1999). *Changes in reading practicum accountability: Preservice teachers are held responsible for children's progress*. Hilton Head, SC: Paper presented at the Annual Meeting of the College Reading Association. (ERIC Document Reproduction Services No. ED447459)	Gr 2,5	Preservice teachers	▪ Not a Randomized Field Trial; pre-post comparison with no control group
29	Ross, S. M., Morrison, G. R., Smith, L. J., & Cleveland, P. (1990). An evaluation of alternative distance tutoring models for at-risk elementary school children. *Computers in Human Behavior, 6*(3), 247-259.	Gr 6	Graduate students	▪ Not a study of a face-to-face tutoring program (compares a cyber-tutoring program to one that also incorporated teleconferencing) ▪ No relevant outcome measure; focus on usage and attitudes

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
30	Taylor, J., & Cox, B. D. (1997). Microgenic analysis of group-based solution of complex two-step mathematical word problems by fourth graders. *Journal of the Learning Sciences, 6*(2), 183-226.	Gr 4	Teachers and a psychologist	▪ Tutors were not volunteers; tutors were teachers and a psychologist
31	Tucker, C. M., Herman, K. C., Reid, A. D., Keefer, N. L., & Vogel, D. L. (1998). The research-based model partnership education program: A 4-year outcome study. *Journal of Research and Development in Education, 32*(1), 32-37.	Gr 3,9	Undergraduates	▪ Not a Randomized Field Trial; control group includes those who did not seek program ▪ No relevant outcomes (only math and reading grades) ▪ No way to separate outcomes for tutees in grade 3 and grade 9
32	Vadasy, P. F. (2002). *Sustainability of promising innovations, November 1, 1998-October 31, 2002: Final Report*. Seattle, WA: Washington Research Inst. (ERIC Document Reproduction Services No. ED 474358)	Gr 1	Community	▪ Not an evaluation of a program; rather, it is a summary of the tutoring treatment with no controls
33	Vadasy, P. F. (2001) Routes to comprehension for second-graders with reading problems: One-to-one tutoring in repeated reading versus comprehension strategy instruction (Doctoral dissertation, University of Washington, 2000). *Dissertation Abstracts International, 61,* 11A.	Gr 2		▪ Not a study with a true comparison group; this study compared two types of treatment
34	Vinograd-Bausell, C. R., & Bausell, R. B. (1987). Home teaching of word recognition skills. *Journal of Research & Development in Education, 20*(3), 57-65.	Gr 1	Parents	▪ Tutoring program did not meet minimum duration of one month (program lasted only 2 weeks)

	Study Name / Brief Program Description	Tutees	Tutors	Reasons for Exclusion
35	Weine, A. M., Kurasaki, K. S., Jason, L. A., & Danner, K. E. (1993). An evaluation of preventative tutoring programs for transfer students. *Child Study Journal, 23*(2), 135-152.	Gr 3-5 (trans-fers)	University undergraduates and parents	▪ Inadequate reporting of statistics; descriptive statistics not reported

13. General References

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*, 237-248.

Devin-Sheehan, L., Feldman, R. S., & Allen, V. L. (1976). Research on children tutoring children: A critical review. *Review of Educational Research, 46*, 355-385.

Elbaum, B., Vaughn, S., Hughes, M.T., & Moody, S.W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*(4), 605-619.

Ellson, D.G. (1976). Tutoring. In N. Gage (Ed.). *The psychology of teaching methods.* Chicago: University of Chicago Press.

Fashola, O.S. (2001). *Building effective afterschool programs.*Thousand Oaks, CA: Corwin Press.

Fitz-Gibbon, C.T. (1977). *An analysis of the literature of cross-age tutoring*, Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED148807).

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goodlad, S. (Ed.). (1995). *Students as tutors and mentors*. London, PA: Kogan Page.

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

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

Michael, B. (1990). *Volunteers in public schools*. Washington, DC: National Academy Press.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Research Council. (1998). *Preventing reading difficulties found in children*. Washington, DC: National Academy Press. Retrieved June 21, 2006, from http://www .readbygrade3.com/prevread.htm#contents

Rekrut, M. D. (1994). Peer and cross-age tutoring: The lessons of research. *Journal of Reading*, *37*(5), 356-362.

Rosenshine, B., & Furst, N. (1969). *The effects of tutoring upon pupil achievement: A research review.* Washington, DC: Office of Education. (ERIC Document Reproduction Service No. ED 064 462).

Rothstein, H. R., Turner, H. M., & Lavenberg, J. G. (2004). *The Campbell Collaboration information retrieval policy brief*. Philadelphia, PA: The Campbell Collaboration.

Shanahan, T. (1998). On the effectiveness and limitations of tutoring in reading. In P. D. Pearson & A. Iran-Nejad (Eds.). *Review of Research in Education* (pp. 217-234). Washington, DC: American Educational Research Association.

Topping, K. J., & Hill, S. V. (1995). University and college students as tutors for schoolchildren: A typology and review of evaluation research. In S. Goodlad (Ed.), *Students as tutors and mentors* (pp. 13-31). London, PA: Kogan Page.

U.S. Department of Education. (1996). *Examples of tutoring programs*. Washington, DC: America Reads Challenge (http:/www.ed.gov/inits/americareads/readexam.htm, Updated August 15, 1996).

U.S. Department of Education. (1997). *Read with me: A guide for student volunteers starting early childhood literacy programs*. Washington, DC: Office of Educational Research and Improvement.

Wasik, B. A. (1998). Volunteer tutoring programs in reading: A review. *Reading Research Quarterly*, *33*, 266-292.

Wasik, B. A., & Slavin, R. A. (1993). Preventing early reading failure with one-to-one tutoring: A review of five great programs. *Reading Research Quarterly*, *28*(2), 179-200.

White House, Office of the Press Secretary. (1997). *President Clinton's America Reads Work-Study Challenge*. Washington, DC: Author.

Wilson, D., & Lipsey, M. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*(4), 413-429.

14. Contact Details for Co-Reviewers

Dr. George Denny
Professor, Educational Statistics
University of Arkansas
234 Graduate Education Building
Fayetteville, AR 72701
Phone: (479) 575-7320
Fax: (479) 575-2492
Email: gdenny@uark.edu

Ginger Albin
Graduate Student, Higher Education
University of Arkansas
234 Graduate Education Building
Fayetteville, AR 72701
Phone: (918) 809-2239
Fax: (479) 575-2492
Email: galbin@uark.edu

Joshua Barnett
Graduate Assistant, Office for Education Policy
University of Arkansas
201 Graduate Education Building
Fayetteville, AR 72701
Phone: (479) 575-3172
Fax: (479) 575-3196
Email: jhb05@uark.edu

Virginia Blankenship
Research Associate, Office for Education Policy
University of Arkansas
201 Graduate Education Building
Fayetteville, AR 72701
Phone: (479) 575-3773
Fax: (479) 575-3196
Email: vhudson@uark.edu

15. Table of Contents