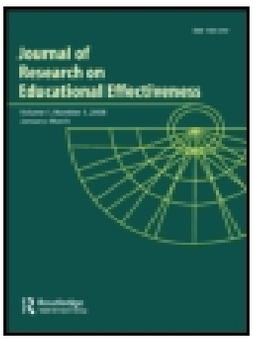


Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

The Impact of Vocabulary Instruction on Passage-Level Comprehension of School-Age Children: A Meta-Analysis

Amy M. Elleman , Endia J. Lindo , Paul Morphy & Donald L. Compton

To cite this article: Amy M. Elleman , Endia J. Lindo , Paul Morphy & Donald L. Compton (2009) The Impact of Vocabulary Instruction on Passage-Level Comprehension of School-Age Children: A Meta-Analysis, Journal of Research on Educational Effectiveness, 2:1, 1-44, DOI: 10.1080/19345740802539200

To link to this article: https://doi.org/10.1080/19345740802539200

Journal of Research on Educational Effectiveness, 2: 1–44, 2009

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345740802539200

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

The Impact of Vocabulary Instruction on Passage-Level Comprehension of School-Age Children: A Meta-Analysis

Amy M. Elleman
Vanderbilt University, Nashville, Tennessee, USA

Endia J. Lindo
Georgia State University, Atlanta, Georgia, USA

Paul Morphy and Donald L. Compton
Vanderbilt University, Nashville, Tennessee, USA

Abstract: A meta-analysis of vocabulary interventions in grades pre-K to 12 was conducted with 37 studies to better understand the impact of vocabulary on comprehension. Vocabulary instruction was found to be effective at increasing students' ability to comprehend text with custom measures ($d = 0.50$), but was less effective for standardized measures ($d = 0.10$). When considering only custom measures, and controlling for method variables, students with reading difficulties ($d = 1.23$) benefited more than three times as much as students without reading problems ($d = 0.39$) on comprehension measures. Gains on vocabulary measures, however, were comparable across reading ability. In addition, the correlation of vocabulary and comprehension effects from studies reporting both outcomes was modest ($r = .43$).

Keywords: Vocabulary, reading comprehension, meta-analysis, instruction, reading difficulties

The ability to understand and gain knowledge from text is a fundamental skill required in every school subject as well as everyday life. Unfortunately, large numbers of school-age children experience significant problems in learning to

Address correspondence to Amy M. Elleman, Peabody College of Education and Human Development, Vanderbilt University, Department of Special Education, Box 228, Nashville, TN 37203, USA. E-mail: amy.m.elleman@vanderbilt.edu

read. In 2007, the National Center for Educational Statistics reported that 69% of eighth-grade students performed below the proficient level in reading based on the National Assessment of Educational Progress (Lee, Grigg, & Donahue, 2007). As students progress through school, the demands of independently extracting and retaining information from text increases. If we are to impact students' ability to independently gain knowledge from text, we must better understand what types of interventions are most effective at increasing students' ability to comprehend what they are reading. One promising area of intervention research is vocabulary instruction. Large individual differences in vocabulary size exist early in school. By the end of second grade, disadvantaged students can lag 2 years behind the average students in their class and 4 years behind students in the upper quartile (Biemiller, 2005).

Both correlational and experimental studies have demonstrated a strong relationship between vocabulary and reading comprehension (Carroll, 1993; Davis, 1942, 1968; Freebody & Anderson, 1983; Marks, Doctorow, & Wittrock, 1974; Spearitt, 1972; Wittrock, Marks, & Doctorow, 1975). Five hypotheses have been put forth to explain the causal link between vocabulary and comprehension (see Nagy, 2005; Stahl & Nagy, 2006). Anderson and Freebody (1981) described three of the five possibilities—instrumentalist, knowledge, and aptitude hypotheses. The instrumentalist hypothesis proposes that knowing vocabulary directly impacts comprehension. Vocabulary training studies, however, have reported variable effects between vocabulary and comprehension. Many studies have shown that vocabulary training impacts comprehension (e.g., Beck, Perfetti, & McKeown, 1982; Kameenui, Carnine, & Freschi, 1982; Stahl, 1983; Stahl & Fairbanks, 1986), whereas others have not found such effects (e.g., Pany & Jenkins, 1978; Tuinman & Brady, 1974; Wixson, 1986).

The knowledge hypothesis states that words are part of larger knowledge structures and that it is these knowledge structures, not the words per se, that impact a person's comprehension. Within this framework, vocabulary knowledge could be considered a proxy for a person's background knowledge. A person that has more knowledge of a subject is likely to better comprehend text about that subject, as well as know more words related to the topic. In contrast, the aptitude hypothesis postulates that there is no causal relationship between vocabulary and comprehension. Proponents of this hypothesis suggest it is a third factor, such as verbal intelligence (see Sternberg & Powell, 1983) or metalinguistic awareness (see Nagy, 2005), that influences both comprehension and levels of vocabulary knowledge. That is, people who are good at making inferences or understanding language are also more likely to be good at learning and using vocabulary.

In addition to the hypotheses discussed by Anderson and Freebody (1981), Mezynski (1983) proposed the access hypothesis. According to this view, comprehension is affected by knowing not only the words in the text but how accurately and quickly those word meanings are retrieved from memory. This

hypothesis is in line with Perfetti's (1985) verbal efficiency theory, which suggests that comprehension is impaired when lower level processes, including the retrieval of word meanings, is not automatic. If word meanings are accurately represented, they will be accessed quickly, freeing up the higher level cognitive processes required for comprehension.

The most recent addition to explain the relationship between vocabulary and comprehension is the reciprocal hypothesis. It is estimated that children learn between 2,000 and 3,000 new words per year (see Stahl & Nagy, 2006). Most of children's growth in vocabulary occurs incidentally, not through instruction or conversation. This phenomenal growth has been attributed to children's exposure to text, because vocabulary in oral speech and other forms of media pale in comparison to the vocabulary found in text (Hayes & Arhens, 1988; Stanovich & Cunningham, 1993). This learning occurs incrementally over time through multiple exposures of words in varied contexts. With each successive encounter, children gain a deeper understanding of a word's meaning (Stanovich & Cunningham, 1993). With a deeper understanding of words and expanded vocabulary, children are better able to understand what they read which leads to increases in text exposure. In this way, vocabulary and comprehension have a reciprocal causal relationship as exemplified in Stanovich's (1986) "Matthew Effects."

Each of the hypotheses is a viable explanation for the relationship between vocabulary and comprehension. However, they are not mutually exclusive, and each probably provides a partial explanation. Although a clear-cut answer explaining this complex relationship is unlikely, understanding the relative contributions of each of these hypotheses is important if we are to design efficient vocabulary interventions that will impact children's comprehension.

One way to better understand the relationship between vocabulary and comprehension is to evaluate the effects of vocabulary interventions designed to increase comprehension. Stahl and Fairbanks (1986) conducted such a review in a meta-analysis of vocabulary interventions conducted in Grade 2 through college. They found an overall mean effect size of 0.97 for comprehension outcomes developed by the researcher and 0.30 for global measures of comprehension. That vocabulary instruction would transfer, even in a limited way, to global measures that were unlikely to contain the taught words was considered an important finding supporting the robustness of certain types of vocabulary instruction (see McKeown & Beck, 2006). Stahl and Fairbanks also found larger effect sizes were associated with (a) activities requiring more depth of processing, (b) contextual information paired with definitional information (vs. either alone), and (c) the type and number of word exposures. Yet the results should be viewed with caution because of several limitations with the meta-analysis. First, Stahl and Fairbanks included effect sizes derived from one-group, pre–posttest designs with no control (e.g., Barrett & Graves, 1981), as well as those derived from pre–posttest, control group designs, even though these two types of effect sizes are not comparable. Second, the authors did not

weight for sample size. This means that all studies, regardless of the sample size, contributed equally to the estimation of the overall effect size. Third, they used multiple effect sizes from each study and did not adjust for the statistical dependency produced by comparing different treatments to the same control. Fourth, they used the control group variance instead of the pooled variance of treatment and control groups to compute effect sizes. Each of these limitations has the potential to distort estimates of the true population effect. Therefore, some of the findings of Stahl and Fairbanks may reflect methodological issues, not effects because of the instructional characteristics of the interventions. Finally, although Stahl and Fairbanks attempted to control for methodological (e.g., measures, control group) and treatment factors through a combination of blocking and stratification, their analysis largely ignored the possibility of interactions among method, participant, and treatment factors.

In addition to the Stahl and Fairbanks (1986) study, the National Reading Panel (NRP) (National Institutes of Children's Health and Development [NICHD], 2000) conducted a narrative review of all published experimental and quasi-experimental studies evaluating vocabulary instruction reported prior to 2000 and concluded that vocabulary instruction is generally effective. The NRP provided the following instructional recommendations based on results from vocabulary interventions in the review: (a) provide direct instruction, (b) supply repetition and multiple exposures in rich contexts, (c) restructure tasks for low-achieving or at-risk students, (d) present activities that actively engage, and (e) employ a variety of instructional methods for optimal results. The NRP also found evidence that the effects of vocabulary instruction varies based on participant characteristics and called for future research to help determine effective interventions tailored for various ages and abilities (NICHD, 2000).

The NRP also raised questions about how researchers should be measuring gains in vocabulary. The panel recommended that more sensitive measures created specifically for the intervention should be used, so that effects could be detected (NICHD, 2000). In contrast to Stahl and Fairbanks (1986), the NRP found only two studies that demonstrated gains on standardized measures. Although informative, the NRP's narrative review has its limitations in informing practice and research. First, the NRP excluded studies based exclusively on students with learning disabilities. Second, the NRP determined an intervention's effectiveness based on whether the results were statistically significant, an approach that can be misleading. Low statistical power because of a small sample, not the intervention's effectiveness, could be the reason a study showed nonsignificant results (Lipsey & Wilson, 2001).

Meta-analysis is a more sensitive and precise approach for detecting differences in study characteristics across studies than narrative reviews (Lipsey & Wilson, 2001). The NRP stated that a meta-analysis could not be conducted concerning vocabulary, because the studies were too varied and there were not enough studies to conduct separate analyses for each of the different types of instruction. However, we felt that it was possible to conduct a meta-analysis by

narrowing the focus of the review to passage-level comprehension outcomes, as this would naturally reduce the variability of types of intervention and outcome measures.

This meta-analysis expands the current literature concerning the impact of vocabulary instruction on comprehension in two ways. First, it is the only review using a meta-analytic approach that focuses on effects of passage-level comprehension because of vocabulary instruction for school-age children. Second, it updates the 1986 findings of Stahl and Fairbanks's meta-analysis of comprehension outcomes using current meta-analytic procedures. We asked the following questions concerning comprehension outcomes for students in grades pre-K through 12:

1. Does vocabulary instruction impact passage-level comprehension? If so, which participant and intervention characteristics are associated with effect size?
2. What methodological characteristics are associated with effect size and need to be controlled to avoid confounding of the findings?
3. Do the same factors that impact comprehension influence vocabulary gains in the same way?
4. Are the effects in vocabulary associated with the effects in comprehension?

METHOD

Study Inclusion Criteria

General Study Characteristics. In an attempt to obtain the fullest body of literature, journal articles, dissertations, reports, and conference papers were eligible for the review. Reports published between 1950 and 2006 were considered to represent vocabulary interventions that would be viable in today's modern classroom and qualified for inclusion.

Intervention. An instructional method focused on increasing word knowledge had to be provided to students with the goal of increasing word knowledge or comprehension. Only studies testing a method of vocabulary instruction that could potentially be used in a classroom setting were selected. Short experimental studies conducted to understand the nature of vocabulary acquisition in which students' differential performance on tasks under varying environmental conditions or in response to differences in instructional materials presented were not included. Studies that only used repeated readings, read-alouds, or independent reading were also excluded from this review unless the intervention contained an instructional method for teaching vocabulary. Although these studies have been shown to be effective in increasing vocabulary and comprehension, the focus of this review was the impact of vocabulary instruction

on comprehension. In addition, studies that contained components intended to address comprehension, as well as, vocabulary, were excluded so that the facilitative effects of vocabulary on comprehension could be isolated.

Outcome Measures. The focus of this review concerns the effect of vocabulary instruction on passage-level comprehension. Therefore, studies had to report at least one measure of comprehension at the passage level. Passage-level text or stories were considered to be more than four sentences in length. Sentence-level measures were excluded because the intent of this review was to consider the impact of vocabulary instruction on the comprehension of text, not the comprehension of the target words in text. Sentence measures are largely dependent on understanding the vocabulary word to make sense of them. If a study used only a cloze procedure to assess comprehension, it was considered a sentence-level measure and was excluded. However, multiple-choice questions, open-ended, and recall measures were acceptable.

Standardized (both criterion and normative referenced) and experimenter-designed measures were acceptable. Conversely, measures in which participants only needed information regarding the target vocabulary to answer items correctly were considered to be an assessment of vocabulary knowledge, not comprehension. These measures were excluded regardless of the test format. In addition, we felt that it was important to understand the impact of vocabulary interventions on the comprehension of younger children (i.e., pre-K to Grade 2), so listening measures of passage-level text were included. Assessments of preschool children's nontextual comprehension have been shown to be predictive of their reading comprehension in second grade and used reliably (van den Broek et al., 2005). Our last criterion for outcome selection was that the results had to be reported in a quantitative format that allowed calculation of an effect size using the standardized difference between means.

Participants and Settings. Included reports involved students in grades pre-K through 12 whose first language was English. Eligible studies were conducted in English-speaking countries in which the instruction and materials were in English. Any study including English language learners in more than 20% of its sample was excluded in an effort to avoid the possible confound of differences because of second language use.

Research Design. Both experimental and quasi-experimental designs were eligible for inclusion. Studies included had to employ either a pretest–posttest control group design, posttest control with randomization, or pretest–posttest within-subject design using counterbalanced conditions. In Monte Carlo studies, effect sizes derived from studies using these designs were found comparable to the standardized difference between means derived from experimental and quasi-experimental designs (Dunlap, Cortina, Vaslow, & Burke, 1996). Conversely, one group's pretest–posttest studies were excluded, absent

counterbalanced treatment ordering, because the effect sizes derived from these studies are not comparable to effect sizes derived from studies comparing treatment and control groups. Single-subject designs were also excluded because of issues of comparability and because it is unclear whether the data from these designs meet the required assumptions for employing the necessary statistical procedures.

Control Groups. Studies were included only if they had one of the following types of control groups: a no-treatment group, no treatment with exposure to the reading materials, a classroom mirror with exposure to reading materials, a classroom mirror with definition instruction, or a classroom mirror that included a weaker intervention. The no-treatment groups received treatment as usual in their class with or without exposure to the same reading materials. The classroom mirror groups were those controls that received a treatment by the researcher designed to mirror classroom practice. Those classroom mirror groups that provided instruction consisted of exposure to the reading materials and used a dictionary definition procedure or an associative method of learning the vocabulary. The final type of control, classroom mirror groups using weaker interventions, were included only if the intervention was considered slightly more involved than the dictionary conditions but substantially less involved than the targeted treatment.

Identification and Retrieval of the Reports

A comprehensive search of previous research and bibliographic databases was conducted in an effort to identify and obtain copies of the entire population of empirical research on vocabulary interventions meeting the eligibility criteria. The process for identifying studies began by extending the search used by the NRP (i.e., *vocabulary AND instruction AND reading AND research AND methods* in the ERIC and PsycINFO databases) to include the years 1950 to 2006, which yielded 3,636 citations. Each of these abstracts was considered for inclusion. In addition to the electronic searches, the reference lists of reviews and prior meta-analyses (i.e., Baumann, Kame'enui, & Ash, 2003; Bryant, Goodwin, Bryant, & Higgins, 2003; Fukkink & deGlopper, 1998; Jitendra, Edwards, Sacks, & Jacobson, 2004; Klesius & Searls, 1990; Marmolejo, 1990; Mezynski, 1983; Stahl & Fairbanks, 1986) were examined for additional studies. A total of 305 full articles were obtained, read, and evaluated for inclusion. This resulted in 37 studies that met the eligibility criteria and were included in this review. Many of the studies that were near-misses were so because of lack of a comparable control (e.g., Barrett & Graves, 1981; Lubliner, 2002; Margosein, Pascarella, & Pflaum, 1982), use of a multicomponent intervention combining vocabulary and comprehension instruction (e.g., Beck & McKeown, 2007; Boettcher, 1983; Ruetzel & Hollingsworth, 1988), use of a sentence-level

comprehension measure instead of a passage-level measure (e.g., Askov & Kamm, 1976; Gipe, 1979; Mastropieri, Scruggs, & Fulk, 1990), lack of enough information to compute an effect size (e.g., Bos, Anders, Filip, & Jaffe, 1989; Otterman, 1955), or cases in which the researchers reported that the intervention was not implemented to a degree necessary to fully represent the intervention as it was intended (e.g., Apthorp, 2006).

Coding the Research Reports

All eligible reports were coded for effect size and study characteristics. Reliability of coding was assessed by having a second person code 14 of the 37 articles. All coders were trained doctoral students. Intercoder agreement was determined using percentage agreement (percentage agreement = agreements/agreements + disagreements). Agreement across categories ranged from 72 to 100% with an overall average of 92%. The two coders reconciled all disagreements by reviewing and discussing each of the articles. For any instance in which reconciliation could not be achieved for a category, a third coder made the final decision. However, there were two variables—definitional-contextual scale and type of exposure—which had to be reconsidered and coded a second time due to poor interrater reliability (56% and 68%, respectively). Agreement for the second attempt was 93% for definitional-contextual scale and 96% for type of exposure.

Effect Size Coding

Measurement Type. Standardized and custom measures were considered separately in this review. Standardized measures are unlikely to contain the target vocabulary, whereas custom measures often do. As a result, custom measures are more sensitive to changes in comprehension because of increased vocabulary knowledge (NICHD, 2000; Stahl & Fairbanks, 1986). In addition to the conceptual differences of the measures, we had four studies that used both standardized and custom measures. We were able to include all of the measures and maintain independence by analyzing these measures separately.

Comprehension Outcomes. Each report was coded by first determining whether the comprehension outcome was from a standardized measure or was a custom measure that was modified or created by the researcher. If more than one posttest comprehension measure was reported, then the format of the test was considered. Multiple-choice formats were selected over recall or other open-ended items because this was the type of format used in most of the studies. A few studies (i.e., Kame'enui et al., 1982; Pany & Jenkins, 1978; Pany, Jenkins, & Schreck, 1982) required special handling because of restricted range on the

selected dependent measure. Specifically, effect sizes for Kame'enui et al. on the inference measures in both experiments, and for the multiple-choice test in Pany et al., were excluded from the analysis, whereas Pany and Jenkins had to be removed entirely because the study had no other comprehension measures fitting the inclusion criteria.

Vocabulary Outcomes. Although the focus of this review was on the comprehension outcomes of vocabulary instruction, we were also interested in the relationship between the vocabulary outcome and comprehension gains. Although we did not require studies to report a vocabulary outcome for inclusion, in the cases in which one was reported we computed effect sizes for the vocabulary measures using the same decision hierarchy as we did for the comprehension measures.

Condition Selection. One of the problems with the prior meta-analysis on the impact of vocabulary on comprehension was the inclusion of multiple conditions using a common control. To avoid this type of data dependency, an independent set of effect sizes was created by choosing one condition per study. The condition that contained the most instructional elements was considered the most intensive treatment and was compared to the least intensive control.

Calculating Effect Sizes. If an author reported an effect size, the effect size was retained without change if it was identified as the *d* statistic. In other cases, we calculated effect sizes by taking the difference between the intervention group and the control group means and dividing by the pooled standard deviations of the means. The *d* statistic presents a ratio of the difference between means to the variance around the means of each group. Whenever possible, *d* was calculated after adjusting for any mean differences at pretest or by calculating the covariate adjusted mean difference and standardizing the difference statistic using the posttest standard deviation (What Works Clearinghouse, 2007). In instances where standard deviations were not reported, they were estimated from reported *t* statistics (see Shadish, Robinson, & Congxiao, 1999; Smith, Glass, & Miller, 1980) or residual sums of squares. For covariate or complex factorial designs, standard deviations were estimated from the sums of squares, which were reconstituted so that sums of squares from factors other than the group factor were recombined with the residual sum of squares. The reestimated residual variance was apportioned equally to treatment and control groups to estimate *d*.

Four studies employed within-subjects, counterbalanced designs (i.e., Pany et al., 1982; Roser & Juel, 1982; Thomas, 1998; Stahl, 1983). Comparisons from these studies were identified following general rules for selecting treatment and control conditions. Although these studies used a correlated design, providing all treatments to all participants, their use of counterbalancing controlled for

ordering, permitting estimations comparable to those from between-subjects designs using matching and randomization (Dunlap et al., 1996).

Although all the aforementioned effects estimates were adjusted or used equivalent comparisons, resulting in better precision and smaller effects, 13 studies reported posttreatment data only and could not be adjusted for possible nonequivalence. Of these 13 studies, 3 (i.e., Beck et al., 1982; McKeown, Beck, Omanson, & Perfetti, 1983; McKeown, Beck Omanson, & Pople, 1985) reported means without standard deviations, forcing a conservative estimate of effects using the exact p values reported from appropriate t or F tests.

Some studies reported outcomes for multiple breakout groups such as male and female or provided separate measures for explicit and implicit comprehension (e.g., Nash & Snowling, 2006; Wixson, 1986). We did not have enough studies reporting these outcomes to consider them separately, so these groups were combined to compute an aggregated effect size using a procedure attributed to Nouri and Greenberg (Cortina & Nouri, 2000). Yet separate effect sizes within studies were retained, if reported across grades, because of our interest in developmental differences. Grade-level effects were considered independently when studies reported both treatment and control groups for each grade level.

Moderator Coding

The differential effects of individual vocabulary studies may be because of systematic differences related to the interventions, designs, participants, settings, and/or general study characteristics. Moderator variables were, therefore, coded in an attempt to account for systematic differences in the research reports. Theory and past research concerning vocabulary guided the selection of potential moderator variables to code from the study reports.

Methodological Characteristics. Methodological characteristics for each study were coded, so differences due to how studies were designed could be ruled out as plausible explanations for the effects or controlled for in the statistical analysis. Minimum standards of study quality were adopted in an effort to consider as much evidence as possible in evaluating the effects of vocabulary on comprehension. Given this, information regarding study quality was coded so that any effects associated with quality indicators could be examined and controlled for, if needed. In addition to coding whether a study randomly assigned students to treatment and control, we coded information about the monitoring of the intervention, the reporting of implementation problems, training for instructors, and information about the reliability of the dependent measures.

Participant Characteristics. As important as the methodological characteristics are, other factors, such as sample characteristics, can also have a significant

influence on a study's outcomes. Yet prior reviews have not comprehensively and systematically considered the interaction of student characteristics and treatment gains in vocabulary. Stahl and Fairbanks (1986) failed to use a moderated analysis that included participant variables, and the NRP (NICHD, 2000) did not include exclusive samples of student with learning disabilities. Students of different ability and grade levels are likely to respond to vocabulary instruction quite differently, so information about the grade level and reading ability were coded when available from the included reports. Student risk for reading difficulty was coded as no identified risk or at risk for a reading problem (i.e., students who were identified as having low scores on a reading test or identified as having a reading problem or learning disability). Ethnicity, gender, and socioeconomic status (SES) were also coded to better understand how these variables affect treatment outcomes.

Intervention Characteristics. Although many methodological and participant characteristics were examined, most of the coded information focused on characteristics of the intervention. Although we coded the type of intervention the authors reported, we decided to use characteristics of interventions for analysis. This decision was made because most of the interventions included multiple components instead of a single treatment, and often interventions with the same elements were assigned different names. These components were based on definitions used by the NRP (NICHD, 2000). Some of the categories of the NRP were deleted because they were not represented in this review due to inclusion criteria, and some categories have been combined to represent studies considered to be similar. In addition to characteristics of the method of instruction, contextual intervention factors such as total hours of treatment, who delivered the treatment (researcher or teacher), number of minutes of instruction per target word, and group size (self-administered, one-to-one, small group, whole classroom) were coded. Other intervention characteristics were coded based on Stahl and Fairbanks's (1986) conceptualizations including depth of processing, definitional-contextual scale, and type of exposure.

Depth of Processing. Stahl and Fairbanks (1986) used a "depth of processing" framework to determine why some interventions might be more effective than others. This framework combined the constructs of the amount of semantic processing and mental effort required to complete an activity. Tasks requiring more semantic processing and metal effort have been shown to produce better recall. Stahl and Fairbanks found that this scale approached, but did not reach, statistical significance. Depth of processing was coded using three categories:

1. Association. This instruction paired association of the new word with its definition or synonym.

2. Comprehension. This instruction required that the student demonstrate comprehension of the meaning of the word by doing something with the definitional information such as classifying terms or providing antonyms.
3. Generation. This instruction required the students to generate a novel oral or written response using the word. Activities included restating of the definition in the student's words or writing an original sentence containing the word.

Definitional-Contextual Scale. Stahl and Fairbanks (1986) reported that balanced programs produced better passage comprehension gains than programs focused on definitional or contextual learning. This scale was coded as the following:

1. Definitional only. Definitions or synonyms are provided without any use of context.
2. Definitional emphasis. A limited amount of context is provided, but an emphasis is placed on learning the definition.
3. Balanced. The instruction contained nearly equal emphasis on definitional and contextual information.
4. Contextual emphasis. Although definitions were provided or derived from context, the focus of this instruction was on understanding the words in context.
5. Context only. The focus of this instruction was to expose the students to words in context with no definitions provided.

Type of Exposure. As discussed earlier, students learn words naturally through repeated exposures in multiple contexts over time. Stahl and Fairbanks (1986) found that using multiple exposures across multiple contexts was more effective than associative or other instructional techniques for producing gains in comprehension. However, they also noted that no firm conclusions could be drawn because only a few studies used associative techniques to produce gains at the passage level. Type of exposure was coded as (a) less than three repetitions in a single context, (b) more than three repetitions in a single context, and (c) more than three repetitions in multiple contexts.

Levels of Discussion. Interventions with higher levels of discussion can be speculated to build on background knowledge and to present the words in multiple contexts, thereby facilitating comprehension gains according to the knowledge and access hypotheses. This variable was coded as (a) little to no discussion or (b) high levels of discussion.

Word Specific Versus Generative Instruction. We were also interested in understanding the differences in comprehension outcomes of interventions focused on teaching specific words and those focused on strategies to increase students'

generative word knowledge. Interventions focused on generative word knowledge were considered activities that taught students how to gain meaning from words beyond the target words taught. For instance, instruction that focused on word parts and instruction that focused on gaining word meaning from contextual clues would be expected to generalize beyond the target words presented in the study, whereas instruction focused on specific words would not be expected to generalize to untaught words. Nagy (2005) discussed the role of metalinguistic skills in vocabulary and comprehension within the context of the verbal aptitude hypothesis. If metalinguistic skill is one of the links between vocabulary and comprehension, then interventions designed to increase students' use of contextual strategies and morphological awareness should increase students' vocabulary and comprehension.

Text Variables. Another variable that may influence the impact of vocabulary on comprehension is the type of text used and tested in a study. Expository texts often contain a higher proportion of context-specific words that are salient to the content presented and are therefore considered more difficult than narrative texts (see Gardner, 2004; Graesser, Golding, & Long, 1991; Wolfe, 2005).

Statistical Procedures

An examination of the distribution of effect and sample sizes was conducted to determine if there were any outliers that could distort results. Outliers were identified based on either effect size (d) or sample size distributions using Tukey's (1977) definition of an extreme outlier as falling three times the interquartile range above the 75th percentile or below the 25th percentile of the distribution. To limit the influence of such effects on subsequent analyses, outliers were Winsorized as recommended by Lipsey and Wilson (2001) to maximum or minimum values at the respective outer limits. Based on this definition, no comprehension effects qualified as outliers, but one study had an unusually large effect in vocabulary ($d = 4.04$). This effect was trimmed to the upper limit value for vocabulary ($d = 2.28$; i.e., Jones, 1984). Similarly, one large sample size ($N = 337$) was reduced ($N = 220$; i.e., Hogan, 1961) to avoid overly influencing the vocabulary and comprehension outcomes.

Effect sizes derived from small samples are known to be biased, so our next step was to adjust the effect sizes using a small sample correction, $1 - (3/4n - 9)$, where n is the total sample size for computing each effect (Hedges, 1982). Each Hedges'sg effect size was then weighted by the inverse of its error variance, $1/SE^2$, to take its proportionate reliability into account (Shadish & Haddock, 1994). A test of homogeneity using the Q-statistic was then applied to establish whether there was more variability in the effects than would be

expected by subject-level sampling error alone (Cochran, 1954; Hedges & Olkin, 1985). The Q-statistic is calculated as

$$Q = \sum w_i \times (ES_i - \overline{ES}),^2$$

in which $w[F1]$ is the inverse variance weight for each effect size i and ES_i is the weighted mean effect size for each i and ES is the weighted mean effect size over all cases of i.This statistic is distributed as chi-square with $k - 1$ degrees of freedom in which k is the number of effect sizes and, when significant, warrants rejection of the null hypothesis that variance in effects are explained by sampling error alone (Lipsey & Wilson, 2001).

For outcomes whose variance exceeded that predicted by sampling error alone (i.e. $p < .05$), mixed-weight regression analyses were first conducted to estimate the moderating influence of method, participant, and intervention variables on comprehension outcomes (see Raudenbush, 1994). Second, a parallel analysis was completed for identifying potential sources of variance in vocabulary effects. Next, we provide an example using a subset of instructional variables to illustrate the importance of testing a conditional model instead of using a stratified analysis. Finally, to better understand the relationship between vocabulary and comprehension outcomes within studies, correlations were computed for the subset of studies reporting both outcome measures.

Table 1. Methodological characteristics of studies by measure type

Treatment monitored				
Yes	5	31.3	6	21.4
Not reported	11	68.8	22	78.6
Type of measure				
Open-ended	2	12.5	7	25.0
Multiple choice	10	62.5	17	60.7
Not reported	4	25.0	4	14.3
Measure reliability				
Reported	3	18.8	7	25.0
Not reported	13	81.3	21	75.0
	0	0.0	10	35.7
Text type				
Narrative				
Expository	2	12.5	12	42.9
Both	4	25.0	2	7.1
Not reported	10	62.5	4	14.3

Note. Three studies have both standard and custom measures.

RESULTS

Descriptive Characteristics of Studies

The literature searches yielded 37 eligible studies from which 44 effect sizes were derived for comprehension outcomes. Of the eligible studies, 28 also administered a vocabulary measure from which 37 independent effect sizes were derived. Across these studies, there were 3,063 participants. To create an

Table 2. Intervention characteristics by measure type

	Standard		Custom	
Characteristic	*N*	%	*N*	%
Processing depth	3	18.8	6	21.4
Associative				
Comprehension	13	81.3	14	50.0
Novel response	0	0.0	8	28.6
Treatment focus				
Word	6	37.5	20	71.4
Generative	10	62.5	8	28.6
Discussion				
None to some	14	87.5	17	60.8
High	2	12.5	11	39.2
Training				
Yes	4	25.0	10	35.7
Not reported	12	75.0	18	64.3
Instructor				
Teacher	8	50.0	9	32.1
Researcher	6	37.5	13	46.4
Self-directed	1	6.3	1	3.6
Cannot tell	1	6.3	2	7.1
Intervention hours				
1–5	1	6.3	16	57.1
6–10	4	25.0	5	17.9
11–15	2	12.5	0	0.0
16–20	1	6.3	3	10.7
21+	7	43.8	2	7.1
Not reported	1	6.3	2	7.1
Targeted words				
5–15	1	6.3	11	39.3
16–25	0	0	5	17.9
26–35	0	0	3	18.8
36+	2	12.5	6	21.4
Not reported	11	68.8	3	10.7

Note. Three studies have both standard and custom measures.

independent set of effect sizes and include as many studies as possible, we split the analysis by studies using custom or standardized comprehension measures. Across measurement types, the typical study used a no-treatment control, used whole class instruction, and was conducted in Grades 3 to 5 with students having no identified risk of reading difficulty. Most studies utilized a multiple-choice format for both comprehension and vocabulary dependent measures. The majority of studies were conducted in 10 or fewer hours. Studies using a standardized measure, in general, had longer interventions that contained more target words and focused more on increasing generative word knowledge than studies using a custom measure (see Tables 1–3).

Across studies, information concerning some methodological, participant, and intervention characteristics was not reported. The majority of studies did not provide information about the reliability of their measure, fidelity of the treatment, or the training procedures for those implementing the intervention. Information about the number of targeted words was also often missing, making it impossible to calculate any type of instructional efficiency score based on number of minutes per instructed word. Likewise, information regarding participants (i.e., SES, ethnicity, and gender) was frequently omitted and therefore could not be considered in the analyses.

After our analysis, it was apparent that effects from standardized measures were minimal (described in detail later in this section) and that effects associated with having reading difficulties were larger than those with no reading problems. We therefore considered the characteristics for studies using custom measures separated by students with reading problems and those without (see Table 4). Only few differences stood out between the studies. Studies

Table 3. Participant characteristics by measure type

	Standard		Custom	
Characteristic	*N*	%	*N*	%
Reading problem				
None identified	14	87.5	23	82.1
Yes	2	12.5	5	17.9
Socioeconomic status				
Low	2	12.5	4	14.3
Middle	3	18.8	13	46.4
High	1	6.3	3	10.7
Not reported	10	62.5	8	28.6
Grade				
Pre-K–2	1	6.3	4	14.3
3–5	9	56.3	17	60.7
6–8	3	18.8	5	17.9
9–12	3	18.8	2	7.1

Note. Three studies have both standard and custom measures.

Table 4. Intervention characteristics by student reading status based on custom measures

	General ability		Reading difficulties	
Characteristics	*N*	%	*N*	%
Type of control				
NT	11	47.8	1	20.0
NT with exposure	4	17.4	0	0.0
CM with exposure	5	21.7	0	0.0
CM with definition	1	4.3	3	60.0
CM with definition +	2	8.7	1	20.0
Instructor				
Teacher	10	43.5	1	20.0
Researcher	9	39.1	4	80.0
Self-directed	1	4.3	0	0.0
Cannot tell	2	13.0	0	0.0
Group format				
One-to-one	3	13.0	0	0.0
Small group	4	17.4	4	80.0
Whole class	16	69.6	1	20.0
Grade				
Pre-K–2	4	17.4	0	0.0
3–5	14	65.2	1	20.0
6–8	3	13.0	2	40.0
9–12	2	4.3	2	40.0
Intervention hours				
1–5	12	52.1	2	40.0
6–10	3	5.9	2	40.0
11–15	2	8.7	0	0.0
16–20	2	8.7	0	0.0
21+	2	8.7	0	0.0
Not reported	2	8.7	1	20.0
Text type				
Narrative	12	52.2	0	0.0
Expository	7	30.4	3	60.0
Both	2	8.7	0	0.0
Not reported	2	8.7	2	40.0
Processing depth				
Associative	5	21.7	1	20.0
Comprehension	11	47.8	3	60.0
Novel response	7	30.4	1	20.0
Treatment focus				
Word	19	82.6	1	20.0
Generative	4	17.4	4	80.0

(Continued on next page)

Table 4. Intervention characteristics by student reading status based on custom measures *(Continued)*

	General ability		Reading difficulties	
Characteristics	*N*	%	*N*	%
Discussion				
No discussion	4	17.4	1	20.0
Moderate	10	43.5	2	40.0
High	9	39.1	2	40.0
Targeted words				
5–15	8	34.8	3	60.0
16–25	5	21.7	0	0.0
26–35	3	13.0	0	0.0
36+	6	26.1	0	0.0
Not reported	1	4.3	2	40.0

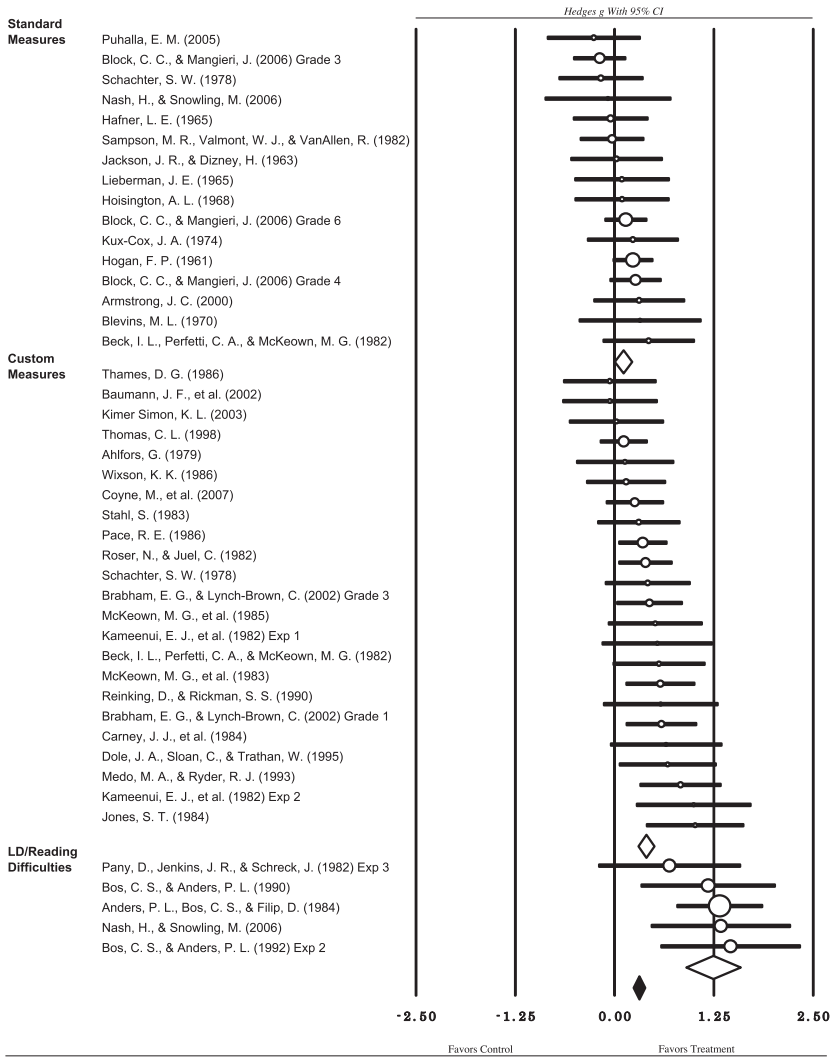
Note. NT = no treatment; CM = classroom mirror.

conducted with students who have reading difficulty were more likely to be short interventions taught in small groups with the researcher as implementer than studies conducted with a universal sample.

The instructional components of each study are shown in Table 5. Some of the components are disproportionately represented when compared across measurement and participant type. For example, four of the five studies conducted with students who had reading difficulties used semantic mapping or semantic feature analysis, and eight of the nine studies using structural analysis employed standardized measures.

Overall Effect Sizes

Comprehension Outcomes. The comprehension effect sizes for standardized measures ranged from –0.26 to 0.43 with an overall random weighted mean effect size of 0.10, which was not significantly different from 0 ($p = .08$). Figure 1 shows the effect sizes for each study and related descriptive information. As expected from prior research (NICHD, 2000; Stahl & Fairbanks, 1986), effects for custom measures were larger than those obtained from standardized measures. The effect sizes for custom measures ranged from –0.06 to 1.46. The overall random-weighted mean effect size was 0.50, and significantly different than 0 ($p < .01$), indicating that students who received vocabulary interventions outperformed students who did not receive such instruction on comprehension outcomes aligned to the treatment.



Note. All effects are residualized for strength of control. Effects for individual studies are represented once within one measure type, but may be represented in both standard and custom measure subgroups. Thus effect estimates are independent within measures but may not be across measures in this display. Symbols are sized proportionate to their estimated precision within each cluster.

Figure 1. Mixed-weight comprehension effects ($K = 44$).

Vocabulary Outcomes. Although standardized measures did not indicate substantial growth for comprehension, standardized vocabulary measures did indicate some improvement in vocabulary knowledge. The effect sizes for standardized vocabulary measures ranged from –0.24 to 0.46 (see Figure 2). The overall random-weighted mean effect size was 0.29 and significantly different than 0 ($p < .01$), indicating that students who received vocabulary instruction

Table 5. Intervention characteristics and effect sizes

	Custom Comprehension Effect Size	Standard Comprehension Effect Size	Custom Vocabulary Effect Size	Standard Vocabulary Effect Size	Standard Comprehension Measure	Control Group Strength	Average Grade	Reading Difficulties/LD	Treatment Hours	Number of Sessions	Number of Target Words	Avg Exposures per Word	Definitional – Contextual Scale	Single or Multiple Contexts	Type of Text	Association Method	Intervention Components: Structural Analysis	Semantic Mapping	Semantic Feature Analysis	Contextual Analysis	Deriving Word Meaning	Elaborative/Rich Instruction	Dictionay/Glossary	Computer/Multimedia	Interactive	Passage Integration	Concept Method	Pre-Instruction	Imagery
Puhalla, E. M. (2005)		−0.26	1.50		SNAP	2	1	+	20.8	5	–	6	BAL	M	MIX	+						+	+		+				
Block, C. C., & Mangieri, J. (2006) Grade 3		−0.19		0.40	Stanford	1	3		36.2	80	–	–	BAL	M	EXP		+												
Thames, D. G. (1986)	−0.06		−0.11			4	2		3.75	3	30	6	CE	S	NAR						+								
Baumann, J. F., et al. (2002)	−0.06		0.57			3	5		10	12	60	5	BAL	M	NAR		+			+									+
Hafner, L. E. (1965)		−0.05		0.10	SRA	1	5		6	12	–	–	BAL	M	EXP					+									
Sampson, M. R., Valmont, W. J., & VanAllen, R. (1982)		−0.03		0.09	Gates	1	3		15	7	–	–	CO	S	EXP					+	+								
Kimer Simon, K. L. (2003)	0.02			0.14		2	PreK		10.7	32	24	12	CE	S	NAR	+		+						+		+	+		
Jackson, J. R., & Dizney, H. (1963)		0.02		0.14	CRCT	1	12		30	35	–	–	BAL	S	EXP	+	+			+	+		+						
Lieberman, J. E. (1965)		0.09		0.46	IOWA	4	5		27	38	95	3	CE	M	MIX	+	+					+				+			
Hoisington, A. L. (1968)		0.09		0.06	Metro	1	6		10	40	–	–	DE	S	EXP		+												
Thomas, C. L. (1998)	0.12		0.36			3	5		15	30	20	10	BAL	M	MIX	+				+			+						
Ahlfors, G. (1979)	0.13		1.12			1	6		5	4	50	5	CE	M	EXP			+											
Block, C. C., & Mangieri, J. (2006) Grade 6		0.14		0.44	Stanford	1	6		36.2	80	–	–	BAL	M	EXP		+												
Wixson, K. K. (1986)	0.15		0.28			4	5		0.75	1	5	7	CE	M	NAR			+								+			
Kux-Cox, J. A. (1974)		0.23		0.21	SRA	3	4		12	24	–	2	CO	S	MIX														
Hogan, F. P. (1961)		0.23	0.43		CRCT	1	11		8.3	50	–	2	DO	S	EXP		+												
Coyne, M., et al. (2007)	0.26		1.43	0.32	Modified SNAP	1	K		18	36	54	–	BAL	M	NAR	+						+			+				
Block, C. C., & Mangieri, J. (2006) Grade 4		0.26		0.37	Stanford	1	4		36	80	–	–	BAL	M	EXP		+												
Stahl, S. (1983)	0.30		2.02			1	5		1	4	10	5	BAL	M	NAR	+									+				
Armstrong, J. C. (2000)	0.31					3	6			30	–	4	DE	S	EXP			+			+		+						
Blevins, M. L. (1970)		0.32		−0.24	Nelson-Denny	1	12		39	13	–	2	DE	S	EXP	+	+						+						
Pace, R. E. (1986)	0.36		0.31			1	10		7.5	150	30	2	BAL	S	EXP														+

Roser, N., & Juel, C. (1982)	0.39		0.31			1	3		0.5	2	8	5	CE	M	NAR	+				+								
Schachter, S. W. (1978)	0.42	−.17	1.26	0.18	Gates	3	5		1.3	7	42	2	DE	S	MIX												+	
Brabham, E. G., & Lynch-Brown, C. (2002) Grade 3	0.44		0.76			3	3		1.5	3	20	–	CE	S	NAR										+			
McKeown, M. G., et al. (1985)	0.51		0.48			1	4		6		24	8	BAL	M	EXP						+							
Kameenui, E. J., et al. (1982) Exp 1	0.54					1	5			1	6	8	BAL	M	NAR	+			+						+	+	+	+
Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982)	0.56	0.43		0.42	IOWA	1	4		37.5	75	104	33	BAL	M	EXP						+				+		+	
McKeown, M. G., et al. (1983)	0.58		0.58			1	4		37.5	75	104	24	BAL	M	EXP					+								
Reinking, D., & Rickman, S. S. (1990)	0.58		1.45			3	6		1	1	32	2	DE	S	EXP	+							+	+				
Brabham, E. G., & Lynch-Brown, C. (2002) Grade 1	0.59		1.01			3	1		1.5	3	20	–	CE	M	NAR										+			
Carney, J. J., et al. (1984)	0.65					1	5		2	2	12	4	BAL	S	EXP	+			+								+	+
Dole, J. A., Sloan, C., & Trathan, W. (1995)	0.67		0.51			4	10		20	20	–	4	CE	S	NAR				+		+						+	+
Pany, D., Jenkins, J. R., & Schreck, J. (1982) Exp 3	0.69					1	4	+	2	3	12	3	DE	S	EXP	+											+	
Medo, M. A., & Ryder, R. J. (1993)	0.83					1	8		2	2	8	5	DE	M	EXP		+		+								+	+
Kameenui, E. J., et al. (1982) Exp 2	0.99					1	5			1	6	5	BAL	M	NAR	+			+						+	+	+	+
Jones, S. T. (1984)	1.01		2.28			2	5		1.5	3	10	9	BAL	M	EXP		+											+
Bos, C. S., & Anders, P. L. (1990)	1.18		0.64			4	7	+	5	6	–	3	CE	M	EXP		+	+	+									
Anders, P. L., Bos, C. S., & Filip, D. (1984)	1.32					4	10	+	2	2	10	5	CE	M	EXP		+		+									
Nash, H., & Snowling, M. (2006)	1.34	−0.08	0.29		NARA II	4	3	+	6	12	12	–	CE	M	EXP	+	+		+	+								
Bos, C. S., & Anders, P. L. (1992) Exp 2	1.46		0.80			4	7	+	2.5	3	–	6	CE	M	EXP		+	+							+			

Note. Pre-K = Pre-Kindergarten, K = Kindergarten; Control Strength (1 = no treatment/no exposure; 2 = no treatment, exposure, no instruction; 3 = treatment mirror, independent with exposure; 4 = treatment mirror with definitional instruction; SNAP = Strong Narrative Assessment Procedure; Stanford = Stanford 9 Achievement Test Battery; SRA = SRA Achievement Series Test; CRCT = Cooperative Reading Comprehension Test; Nelson-Denny = Nelson Denny Reading Test; ITBS = Iowa Test of Basic Skills; Metro = Metropolitan Achievement Test; NARA II = Neale Analysis of Reading Ability; DO = Definition only; DE = Definitional emphasis; BAL = Balanced; CE = Context emphasis; CO = Context only; S = Single; M = Multiple; NAR = Narrative; EXP = Expository.

Table 6. Zero-order correlations of selected method variables with custom measures effects for comprehension ($N = 28$) and vocabulary ($N = 22$)

Method variable	r_{comp}	r_{vocab}
Control group strength	.13	−.30
Experiment vs. quasi-experiment	.03	−.38

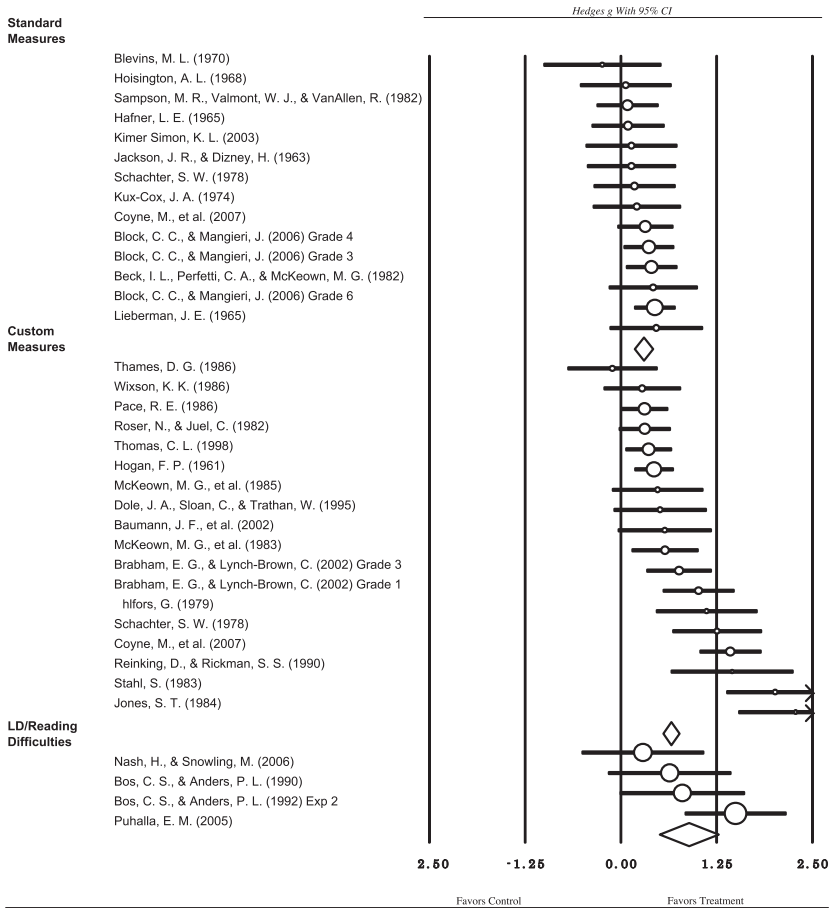
Note. Random-weighted analysis; control group, 1 (*no treatment, no exposure or exposure with no treatment*), 2 (*treatment mirror with exposure*), 3 (*treatment mirror with definition instruction*), 4 (*definition instruction plus added instruction*).

increased their word knowledge on standardized tests. As expected, vocabulary outcomes for custom measures showed the largest effects. The mean effect sizes for custom measures of vocabulary ranged from –0.11 to 2.28. The overall random-weighted effect size was 0.79 and significantly different than 0 ($p <$ .01), indicating that students who received vocabulary instruction knew more vocabulary words than students in control conditions.

Analysis of Moderator Effects

Comprehension Outcomes. For standardized measures of comprehension, the Q-test was not significant, $Q_{resid}(16) = 9.61$, $p = .84$, suggesting that these effect sizes were homogenous. There was little variance that could be attributed beyond sampling error, so no moderator analysis was conducted for these effect sizes. In contrast to the standardized measures, the Q-test from custom measures was significant, $Q_d\ (28) = 46.88$, $p < .01$ indicating excess variance and meriting a moderator analysis to identify study characteristics associated with effect size.

First, we examined the methodological characteristics (i.e., type of control group, quasi-experiment) to determine if any were associated with effect size, so active method variables could be controlled in subsequent analyses of substantive variables of interest. Zero-order random effects correlations for the method characteristics with effect size were conducted using maximum likelihood estimation and are shown in the first column in Table 6. Although these correlations helped in determining which method variables had influential relationships with effects, caution must be exercised when interpreting them, because they ignore all other factors moderating treatment effects including student intervention and treatment differences. Only control group strength, although not significant, showed a nontrivial relationship with effect size (i.e., $\beta > .10$), and was carried forward as a control moderator in subsequent analyses of participant and treatment characteristics.



Note. All effects are residualized for strength of control. Effects for individual studies are represented once within one measure type, but may be represented in both standard and custom measure subgroups. Thus effect estimates are independent within measures but may not be across measures in this display. Symbols are sized proportionate to their estimated precision within each cluster.

Figure 2. Mixed-weight vocabulary effects ($K = 36$).

Next, a series of inverse-variance weighted random effects multiple regressions were employed, each including a single study characteristic while controlling for control group strength, to identify relationships between each study characteristic and effect size without the confounding influence of other study characteristics (Raudenbush, 1994). This analysis indicated that the year of publication, although not significant, negatively correlated with effect size and that dissertations were more likely to be associated with smaller effects than reports and journal publications (see first column in Table 7).

We next considered student characteristics associated with effect size, including the average grade of the participants in the study and whether the participants were identified as having a reading difficulty. Both grade level

Table 7. Conditional correlations of study characteristics with custom measure effects for comprehension ($N = 28$) and vocabulary ($N = 22$) controlling for method variables

Study characteristic	β_{comp}	β_{vocab}
General		
Year of publication	−.19	.21
Dissertation (1) vs. other (0)	−.40	−.04
Student		
Grade level	.37	−.19
Reading difficulties (1) vs. no identified risk (0)	.73	.16

Note. Mixed-weighted analysis. Each treatment characteristic moderator was entered individually and tested in the presence of method moderators control group, 1 (*exposure with no treatment*), 2 (*classroom mirror with exposure*), 3 (*classroom mirror with definition instruction*), 4 (*definition instruction plus added instruction*) and experiment versus quasi-experiment.

and reading status correlated positively with effect size, with reading status correlating significantly. This pattern held when only considering method and participant characteristics (see first column in Table 8). This indicated that students identified as having reading difficulties benefited more from vocabulary instruction on comprehension outcomes than students who had no indicated risk of a reading problem or disability. Although grade level was not significant, it approached significance. One of the reasons for the lack of significance may be because of the limited range of grades included in the analysis (i.e., 50% of the studies were conducted in Grades 3–5).

When considering control group strength and reading status simultaneously in the full regression model, only reading status remained significant (see Table 9). The amount of variance left unexplained was negligible, thus rendering the model fixed, $Q(25) = 24.60$, $p = .48$. A fixed model indicates that there is no systematic variance associated with other method, student, or intervention characteristics for comprehension outcomes. Although the comprehension effects were very similar, and no moderator analysis was warranted, the descriptive information listed in Table 5 shows a possible pattern for interventions conducted with students not identified as having reading difficulties. Studies with slightly higher effects tended to use preinstruction and imagery.

Vocabulary Outcomes. Although an overall positive effect was found on standardized vocabulary measures ($d = 0.33$), the Q-test was not significant, $Q(14) = 7.00$, $p = .90$, signifying there was no variance other than what would be expected from sampling error. Hence, the model for standardized vocabulary

Table 8. Simultaneous correlations of study characteristics with comprehension ($N = 28$) and vocabulary effects ($N = 22$) from custom measures

Study characteristic	β_{comp}	β_{vocab}
Method		
Control group strength	–.16	–.24
Experiment vs. quasi-experiment	–.06	–.34
Student		
Grade level	.18	–.13
Reading difficulties vs. no identified risk	.68	–.11
Instruction		
Total hours of treatment	—	–.38
Exemplary studies vs. other	—	–.01
Small group vs. whole class	—	.21
High vs. lower discussion level	—	.47
Word specific focus vs. generative focus	—	.13
Narrative vs. expository text	—	–.22

Note. Mixed-weighted analysis. The comprehension model is fixed without instructional variables, $Qresid(23) = 22.97$, unlike vocabulary, $Qresid(17) = 62.11$; control group, 1 (*exposure with no treatment*), 2 (*classroom mirror with exposure*), 3 (*classroom mirror with definition instruction*), 4 (*definition instruction plus added instruction*); discussion level (1 = high, 0 = other).

measures was fixed, justifying no further moderator analysis for this subset of effect sizes.

However, vocabulary effects from custom measures were significant, $Q(22) = 91.54$, $p < .01$, indicating substantial unexplained variance in effects. This variance could be because of systematic differences from the method,

Table 9. Relationships between selected practically significant study characteristics and comprehension effects ($N = 28$)

		95% Confidence interval				
Study characteristic	B weight	Standard error	Lower limit	Upper limit	p	β
Intercept	0.47	.10	.29	.66	<.01	—
Control group strength	−0.04	.04	−.12	.04	.28	−.17
Reading difficulties vs. no identified risk	0.92	.20	.53	1.30	<.01	.73

Note. Mixed-weighted analysis. $Qresid(25) = 24.6$; control group, 1 (*exposure with no treatment*), 2 (*classroom mirror with exposure*), 3 (*classroom mirror with definition instruction*), 4 (*definition instruction plus added instruction*).

participant, and intervention characteristics. Therefore, the same statistical procedures were followed, as with the custom comprehension measures, to detect which study characteristics were associated with effect size. The method characteristics used for comprehension were entered first. The magnitude and direction of these correlations were different from the comprehension effects (see second column in Table 6). Control group strength and experiment versus quasi-experiment were sufficiently correlated with vocabulary effects (i.e., $\beta > .10$) to justify controlling these method variables in further analyses.

Next, general study and student characteristics were entered separately into the regression analysis. Although these were not significant, the year of publication and reading status variables were nontrivial (see column 2 in Table 7). Grade level was negatively correlated with effect size for vocabulary measures but was positively correlated with effect sizes derived from comprehension measures. This suggests that for younger students the benefit of vocabulary instruction is more apparent on measures of vocabulary, whereas for older students the benefit is more apparent on measures of comprehension.

As with the comprehension measures, we were interested in what intervention characteristics were associated with vocabulary effect size. Unlike the analysis with the custom comprehension measures, which became a fixed model after entering reading status, the vocabulary analysis showed that there was unexplained variance left to model. Therefore, intervention characteristics believed to explain differences in the effect sizes of the vocabulary outcomes were individually entered while controlling for control group strength, experimental design, and reading status. We did not have enough degrees of freedom to address all of the instructional variables of interest, so we decided to test the exemplary studies against all the other studies. Current recommendations in vocabulary instruction suggest that the combination of generation and multiple repetitions in multiple contexts would yield the largest effects. The association between intervention characteristics and vocabulary outcomes is shown in the second column in Table 8. Surprisingly, exemplary studies did not have any advantage over other studies for vocabulary outcomes.

At this point in the model development, we still had too few degrees of freedom to address all of the instructional variables of interest simultaneously. Retaining the control group moderator of necessity, we therefore reduced the potential moderators to those that were most correlated with effect size and least correlated with each other. These variables (experimental design, total hours of treatment, group format, and level of discussion) were entered simultaneously in a regression model (Table 10). We then reduced the model to those effects that were statistically significant. Although initially in the analysis smaller groups seemed to produce larger effects, when considered with the other instructional variables, group size was not statistically significant. As expected, studies using more stringent control groups (i.e., control groups provided with the target words or some type of instruction vs. no instruction) had smaller effects. Also, as expected, studies using experimental and within-subject designs produced

Table 10. Relationships between selected practically significant study characteristics and vocabulary effects ($N = 22$).

Study Characteristic		B weight	95% Confidence Interval: Standard error	Lower limit	Upper limit	p	β
Model 1	Intercept	1.24	0.28	0.68	1.79	<.01	—
Method	Control group strength	–0.16	0.06	–0.28	–0.03	.01	–.39
	Experiment vs. quasi-experiment	–0.44	0.19	–0.81	–0.07	.02	–.38
Instruction	Total hours of treatment	–0.01	0.004	–0.02	–0.001	.03	–.37
	Small group vs. whole class	0.22	0.17	–0.11	0.56	.19	.20
	High vs. lower discussion level	0.53	0.17	0.20	0.87	<.01	.48
Model 2	Intercept	1.37	0.28	0.81	1.93	<.01	—
Method	Control group strength	–0.16	0.07	–0.29	–0.03	.02	–.39
	Experiment vs. quasi-experiment	–0.47	0.20	–0.85	–0.09	.02	–.41
Instruction	Total hours of treatment	–0.01	0.00	–0.02	–0.001	.03	–.38
	High vs. lower discussion level	0.51	0.18	0.15	0.86	<.01	–.45

Note. Mixed-weighted analysis. Qresid (17) = 38.8; discussion level (1 = high, 0 = other); control group, 1 (exposure with no treatment), 2 (classroom mirror with exposure), 3 (classroom mirror with definition instruction), 4 (definition instruction plus added instruction).

smaller effects than those using quasi-experimental designs. Surprisingly, this model indicated that shorter studies had better outcomes on vocabulary measures. This finding was counterintuitive, so we took a closer at the variable and the vocabulary studies using custom measures. We found that most of the studies using custom measures in vocabulary were short in duration. More than half the studies were conducted in less than 10 hr, and only three were conducted in 40 hr or more. Longer studies will have to be implemented before any firm conclusions can be made about length of treatment and outcomes for vocabulary. One instructional variable, level of discussion, seemed to be important in vocabulary learning. The use of high levels of discussion was associated with greater effects on custom vocabulary measures.

An Illustrative Example of the Necessity of Using a Conditional Analysis. Although we were unable to consider the instructional variables of depth and exposures in our full analysis because of the limited number of studies, we decided to use these variables to demonstrate the importance of using an analysis that considers the conditionality of the effects. Table 11 shows three different ways to look at the effects. We first considered the unconditional effects stratified by depth and type of exposure for vocabulary and comprehension custom measures. Although these strata are arranged hierarchically, the effect sizes associated with each level do not show the theoretically anticipated pattern, with the exception of comprehension effects associated with type of exposure. Second, we considered the interaction effects for depth of instruction by type of exposure. Overall, the anticipated pattern for exposure is observable within each level of depth for both comprehension and vocabulary. The type of exposure also seems to be important with many repetitions in context producing better results than less than three repetitions in a single context for both vocabulary and comprehension. For instance, there is a marked difference between single and multiple contexts for instruction requiring comprehension of the target words for vocabulary outcomes, and to a lesser extent for comprehension outcomes. Notice, however, that most studies used multiple contexts for comprehension. Only two used multiple repetitions in a single context. On the other hand, studies using an associative task did not attempt to use multiple contexts. Therefore, no clear comparison can be made across levels of depth. Although these patterns are interesting, direct comparison of these studies are tenuous because (a) we do not have studies to represent all levels of each factor and (b) because some of the variance attributed to these categories could be due to methodological and participant factors that have not been taken into account. Although it presents an imperfect solution to this problem, the use of statistical control in a conditional model allows reasonable comparisons of the effects. As can be seen in the final section of the table, once the active method and participant factors are controlled, the differences between categories are reduced. These methods must be considered in any analysis of studies employing different methodological design elements. This example demonstrates the pitfalls of

Table 11. Unconditional and conditional impacts of instructional depth, multiple exposures, and multiple contexts from custom-measures of comprehension ($N = 28$) and vocabulary ($N = 22$) effects

Study Characteristic	Voc ES	*n* voc	Comp ES	*n* comp
Unconditional effects				
Depth				
Associative	0.81	6	0.48	6
Comprehension	0.75	10	0.53	14
Generation	0.86	6	0.48	8
Exposures				
<3 exposures–Single context	0.62	3	0.38	2
Many-single context	0.71	5	0.43	8
Many-multiple contexts	0.88	14	0.55	18
Interaction effects				
Associative				
<3 exposures–Single context	0.57	3	0.38	2
many-single context	1.00	3	0.55	4
Many-multiple contexts	—	0	—	0
Comprehension				
<3 exposures–Single context	—	0	—	0
Many-single context	0.20	2	0.30	2
Many-multiple contexts	0.90	8	0.65	10
Generation				
<3 exposures–Single context	—	0	—	0
Many-single context	—	0	—	0
Many-multiple contexts	0.86	6	0.48	8
Conditional effects[a]				
Associative				
<3 exposures–Single context	0.97	3	0.64	2
Many-single context	0.96	3	0.56	4
Many-multiple contexts	—	0	—	0
Comprehension				
<3 exposures–Single context	—	0	—	0
Many-single context	0.72	2	0.42	4
Many-multiple contexts	1.02	8	0.46	10
Generation				
<3 exposures–Single context	—	0	—	0
Many-single context	—	0	—	0
Many-multiple contexts	1.13	6	0.44	8

Note. Mixed weight unconditional means.

[a]Conditional upon control strength, experimental design, hours of instruction and student reading difficulties; control group, 1 (*exposure with no treatment*), 2 (*classroom mirror with exposure*), 3 (*classroom mirror with definition instruction*), 4 (*definition instruction plus added instruction*).

simple stratification and highlights the need for more studies addressing these specific questions.

Relationship Between Vocabulary and Comprehension. Differences were found in the pattern of effects for vocabulary and comprehension. The overall comprehension effect for students with reading difficulties ($d = 1.23$) was much larger than that for students with no indicated problem ($d = 0.39$). However, for vocabulary outcomes, both groups made similar gains from instruction ($d = 0.84$ for students with no indicated problem and $d = 0.79$ for students with reading difficulties). We also found differences in the amount of variance between the two outcomes with vocabulary having more variance than the comprehension outcomes (see Table 12 and Figure 3).

In additiona, we were interested in how much correspondence existed between gains on the vocabulary measures and the comprehension measures, so we examined those studies reporting both vocabulary and comprehension effects ($N = 20$). A regression analysis was conducted predicting outcomes in comprehension from those in vocabulary. This analysis was stratified for students with reading difficulties and those without, because of expected differences on the comprehension outcome. If we assume the instrumentalist hypothesis to be true we would expect that comprehension effects would be strongly and positively correlated with vocabulary effects. This turns out not to be the case (see Figure 3). Although intercept values differ for the two student types, their slope is the same ($r = .43$), indicating only a modest correlation between the effects in vocabulary and comprehension. This lack of relationship may be partially attributed to the large variance in observed vocabulary effects. In fact, the fixed-weight variance of vocabulary effects ($var = .32$) was fully 3.5 times greater than that of corresponding comprehension effects ($var = .09$) from the same studies. If we assume lower measurement reliability to have contributed to this variance, we can impute a reasonable estimate of test–retest reliability to correct for this attenuation. In this case we imputed a test–retest reliability value for comprehension ($r_{yy'comp} = .80$), and from this imputed estimated vocabulary reliability as a function of the effect variance ratio (i.e., $r_{yy'voc} = .80 \div \sqrt{.32/.09} = .42$). Adjusting for reliability yielded an acceptable correction for attenuation and a larger association estimate,

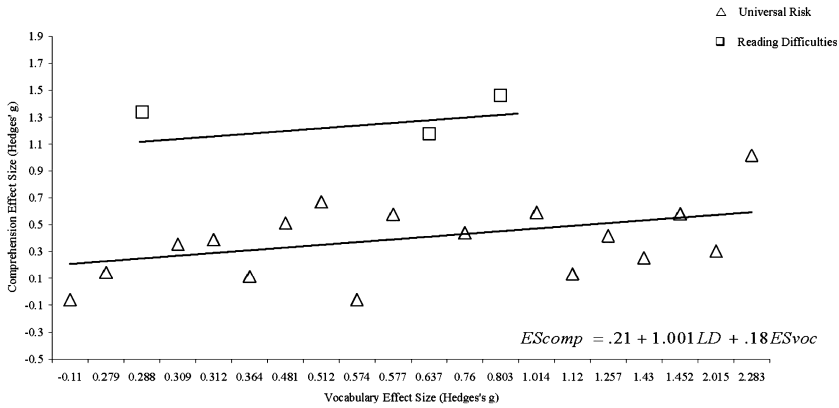
$$r_{voccomp'} = \frac{r_{voccomp}}{\sqrt{r_{yy'comp} \times r_{yy'voc}}} = \frac{.43}{\sqrt{.8 \times .42}} = .74.$$

A corrected correlation of .74 means that roughly 55% of the variance in comprehension gains may be explained by vocabulary growth. This best estimate of the relationship still falls short of what we would expect if vocabulary directly impacted comprehension as suggested by the instrumentalist hypothesis.

Table 12. Effects considered in a serial reduction in unexplained between-study heterogeneity

	Standardized[a] measures					Custom measures all participants					Custom measures no identified risk					Custom measures LD/reading difficulties				
	d	$p_{(d)}$	Q	k	$p_{(Q)}$	d	$p_{(d)}$	Q	k	$p_{(Q)}$	d	$p_{(d)}$	Q	k	$p_{(Q)}$	d	$p_{(d)}$	Q	k	$p_{(Q)}$
Comprehension (N = 44)																				
Fixed	0.10	.08	10.57	16	.78	0.45	<.01	46.88	28	.01	0.39	<.01	23.99	23	.35	1.23	<.01	1.80	5	.77
Random	0.10	.08	10.57	16	.84	0.50	<.01	29.08	28	.36	0.39	<.01	22.11	23	.45	1.23	<.01	1.80	5	.77
Stahl & Fairbanks (1986)	0.30	—	—	15	—	0.97	—	—	40	—	—	—	—	—	—	—	—		-	-
τ^2	<.01	—	—	—	—		0.45	—	—	—	—		<.01	—	—	—	—		<.01	—
H^2	0.70	—	—	—	—	1.74	—	—	—	—	1.09	—	—	—	—	0.45	—			
I^2	<.01	—	—	—	—	0.42	—	—	—	—	0.08	—	—	—	—	<.01	—			
Vocabulary (N = 36)																				
Fixed	0.29	<.01	6.79	14	.91	0.66	<.01	91.54	22	<.01	0.64	<.01	84.53	18	<.01	0.88	<.01	5.72	4	.12
Random	0.29	<.01	6.79	14	.91	0.79	<.01	25.47	22	.23	0.79	<.01	22.55	18	.16	0.84	.002	2.80	4	.42
Stahl & Fairbanks (1986)	0.26	—	—	17	—	1.70	—	—	55	—	—	—	—	—	—	—	—		—	—
τ^2	<.01	—	—	—	—	0.19	—	—	—	—	—	—	—	—	—	0.15	—			
H^2	.49	—	—	—	—	4.36	—	—	—	—	—	—	—	—	—	1.91	—			
I^2	<.01	—	—	—	—	0.77	—	—	—	—	—	—	—	—	—	0.48	—			

[a] No studies used standardized vocabulary measureswith LD/Reading Difficulties.Q *is a* χ-distributed statistic, and like χ tests for a single comparison Ho: $Q_{resid} = N - 1$ *df*. H^2 presents an efficient estimate of between-study variance observed proportionate to that expected (i.e., $H^2 = Q/df$). I^2 corresponds conceptually to the coefficient of alienation (i.e., $1 - r^2$) in presenting the proportion of between-study variance remaining unexplained in the model (Higgins & Thompson, 2002).

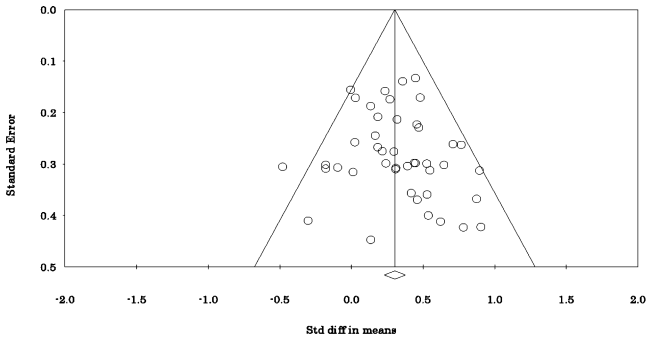


Note. Mixed-weight meta-regression.

Figure 3. Meta-regression of custom-measure comprehension effects onto custom-measure vocabulary effects from studies reporting both effect types ($N = 20$).

Publication Bias

One threat to the conclusions of meta-analysis that must be addressed is small sample, or "publication-bias" (e.g., Duval & Tweedie, 2000; Sterne & Egger, 2001). Upward biasing of the mean population effect estimates may occur because of nonreporting of underpowered studies or studies producing negative treatment effects. Based on the assumption of normal distribution of residuals around the mean, publication bias can be identified by plotting observed effects versus standard error. To ensure that the sample of studies used in this article did not contain a small sample bias, we plotted all of the independent effects



Note. Studies reporting both standard and custom measures of comprehension ($n = 3$) are indicated twice. Effects shown have been residualized for dependent measure (custom vs. standard) and participant type (reading difficulties vs. no identified risk) and plotted as a function of standard error.

Figure 4. Funnel plot of residualized comprehension effect sizes ($N = 44$).

($N = 37$) residualized for variance attributable to measurement (standard vs. custom), reading ability (reading difficulty vs. no reading problem), and method variables (experimental design and control group). Although visual analysis of the funnel plot in Figure 4 identifies one study (Puhalla, 2005) with a larger negative effect than predicted by sampling error, the variation is not excessive. Similarly, Egger's regression intercept value (.469) computed by regressing the ratio of the effect to standard error onto the standard error fell with the 95% confidence interval of expected values ($p = .177$, one-tailed) suggesting no upward biasing in the current sample of effects.

DISCUSSION

The complicated relationship between vocabulary and comprehension has intrigued and eluded researchers over the past century. This meta-analysis was conducted to help clarify some of the theoretical and practical issues concerning the impact of vocabulary training on comprehension. One of the issues surrounding vocabulary has been the degree to which vocabulary training transfers to different types of comprehension tasks. Although a positive overall effect of vocabulary training on comprehension assessed with custom measures was found, the effect for standardized measures was minimal. By contrast, Stahl and Fairbanks reported larger effects for global measures of comprehension. This divergence in findings could be due to differences in inclusion criteria or the methods used to evaluate effects. We found two studies that produced small to moderate gains in comprehension (i.e., Beck et al., 1982; Blevins, 1970). Our finding is similar to that of the NRP who found only two studies produced gains on standardized vocabulary measures. Pearson, Hiebert, and Kamil (2007) offered three possible explanations for the limited impact of vocabulary training on standardized measures: (a) There is no causal link between vocabulary and comprehension, (b) vocabulary instruction does not transfer beyond the taught target words and texts in which it is learned, or (c) existing measures are not sensitive enough to detect changes in comprehension due to vocabulary instruction. The findings of our study provide support for the second hypothesis but do not exclude the possibility that standardized measures could be improved to capture vocabulary growth.

Reading researchers have acknowledged the shortcomings of existing vocabulary and comprehension measures and have called for the creation of reliable, valid, and sensitive measures (see Paris & Stahl, 2005; Pearson et al., 2007). There has been an emphasis on determining the effectiveness of vocabulary instruction based on attaining gains on standardized comprehension tests. Although creating standardized comprehension measures sensitive enough to detect vocabulary growth will greatly improve our understanding of developmental vocabulary growth and, possibly, effects from long-term interventions, it may be unrealistic to consider gains on standardized tests our only benchmark

for determining whether a vocabulary intervention is beneficial. If vocabulary instruction of target words or strategies helps children better understand the local context of what they are reading, it is a worthwhile endeavor. Even if standardized tests are improved and can detect differences due to interventions, it is unlikely that these measures will capture growth from short-term vocabulary interventions. Not only have custom measures been necessary to detect these changes in past studies, they will likely remain important in the future.

The good news is that the overall positive effects found for custom measures suggest that vocabulary training does increase comprehension for all students. In addition, students identified as having reading problems made more than three times the gains than students with no indicated reading problem. This pattern, however, was not the same with the vocabulary outcomes. Students with reading difficulties made equivalent gains in vocabulary knowledge as those without. This finding suggests that vocabulary instruction is more beneficial for understanding text for students with reading problems than for those without reading difficulties.

Two of the hypotheses explaining the relationship between vocabulary and comprehension, access and knowledge, could be used to explain why students with reading problems benefit more from similar levels of vocabulary knowledge. Poor readers are likely to have difficulties with lower level skills such as decoding and quick access to word meanings (Mezynski, 1983). If students learn target words contained in the text, it may free up cognitive resources that can be allocated for higher level processes of integrating text (Mezynski, 1983; Perfetti, 1985). In comparison to students who do not have lower level deficits, poor readers will likely benefit more from learning vocabulary, because they can access words more quickly, thus alleviating cognitive resources and increasing their capacity to engage in the higher level skills required for comprehension. Another reason that students with reading difficulties benefit more from vocabulary instruction on measures of comprehension may be due to increases in knowledge. Students with reading comprehension problems have been shown to have deficits in background knowledge (e.g., McNamara & McDaniel, 2004). All of the interventions with the students who had reading difficulties involved a moderate to high level of discussion, and most were conducted in small groups. It could be that discussion of the target words increased the students' knowledge of the text topics. Therefore, gains in comprehension may be due to increased knowledge of the topics, in addition to the words the students learned.

Although we can recommend vocabulary instruction for increasing comprehension, especially for struggling readers, the not-so-good news is that we cannot provide recommendations about which vocabulary techniques or interventions are best at promoting comprehension. The studies conducted with students who had no reading problems (universal sample) were more varied in their intervention characteristics than the studies conducted with students with reading problems, yet these studies were invariant in their relationship to effect

size. After considering the type of measurement and reading ability, the model became fixed, indicating that there were no other treatment variables that would explain the variation in effect sizes between studies for comprehension. In other words, no matter what type of vocabulary instruction was used, it produced the same effect on comprehension as any other type of vocabulary instruction. This finding is similar to that of Petty, Herold, and Stoll (1968), who found an overall positive effect of vocabulary but could not determine which interventions were most effective. Conversely, Stahl and Fairbanks (1986) found larger effects for studies that required students to understand the words in multiple contexts and to make a novel response using the target words. Our study, however, found no overall differences in intervention characteristics after taking methodological and participant characteristics into account. Our estimations of individual effect sizes for studies included in both reviews were different in many cases. Effect estimates in this study were smaller, which may be a product of our use of single rather than multiple effects and standardization by pooled variance rather than control-group variance.

Although there were no differences that could be detected beyond method, measurement, and participant characteristics for the comprehension measures, there were detectable intervention differences in the vocabulary outcomes. Studies that utilized higher levels of discussion were associated with larger effects for vocabulary outcomes.

We had hoped to consider intervention characteristics that, in addition to informing practice, might help us better understand the relationship between vocabulary and comprehension. The comprehension effects on custom measures for universal students suggest that there are no systematic differences due to intervention characteristics. It could be argued that this finding supports the instrumental hypothesis in which the effects of comprehension are solely due to increased vocabulary knowledge. If this is true and vocabulary is causally related to comprehension, then gains in comprehension should correspond to gains in vocabulary. We decided to test if this was the case in those studies that assessed both comprehension and vocabulary using custom measures. For both students with reading problems and those without, the relationship between vocabulary and comprehension gains was weak, or at least not as strong as it should be if increased vocabulary knowledge is responsible for better comprehension. We therefore have to conclude that either increased vocabulary knowledge does not directly lead to comprehension gains or the measurement of these processes is inadequate for fully capturing the relationship between vocabulary and comprehension. To evaluate the merits of either line of reasoning, we must again visit issues of measurement.

The weak relationship may also be due to the construction of poorly conceptualized, unreliable measures as discussed earlier. Nearly two thirds of the studies included in this analysis did not report any type of reliability for the measures they created, and most provided no conceptual rationale for the way they measured vocabulary or comprehension. Pearson et al. (2007) made a valid

point that we will not be able to adequately explain the relationship between comprehension and vocabulary until we develop and test measures "that are conceptually rich as the phenomenon . . . they are intended to measure" (p. 283). Until we have such measures, we will not be able to determine to what extent vocabulary knowledge contributes to comprehension.

A case could also be made that a third factor such as increased background knowledge, word strategies, or word awareness is responsible for indirectly increasing both comprehension and vocabulary within these studies. Unfortunately, we have no way to test this hypothesis, because few researchers tested or controlled for these possible factors in the studies included in this review.

Limitations

Some of the questions related to participant and intervention characteristics could not be answered because they were not reported. For example, relationships between effect size and participant characteristics such as SES, gender, and race could not be examined. Another issue is that descriptions of interventions and procedures were often insufficient to determine how many target words were taught, how many exposures the students had to the target words, or the types of contexts contained in the training materials or measures. In addition to the lack of reporting for measure reliability, another limitation to the generalization of these findings is that more than two thirds of researchers did not report information on treatment fidelity or training of the intervention implementers. If there is no information on whether the intervention was implemented as it was intended to, we cannot be confident that the results were due to the intervention.

The findings of this study must be considered within the limitations of meta-analysis. Meta-analysis only affords the ability to generalize from the characteristics of existing studies. The distribution of studies across measurement, intervention, and participant characteristics must also be considered when generalizing any findings. For example, the comprehension effects from studies using structural analysis showed poor results. If we only consider this information, we might conclude that instruction in structural analysis of words is ineffective at promoting comprehension. However, the majority of studies testing structural analysis used standardized, not custom measures. This intervention needs to be tested using more sensitive measures before we can determine its effectiveness. Similarly, most of the studies were conducted in Grades 3 to 5. More research needs to be conducted in the early, middle, and high school grades before generalizations can be made concerning the impact of vocabulary interventions across developmental periods. We also need to consider the effects of long-term studies. Biemiller (2005) suggested that to increase general comprehension, a child would have to learn at least 1,000 root words over the primary years. None of the studies reviewed spanned longer

than 1 year, and most of the studies were conducted in less than 15 hr. We need long-term studies across the primary years to truly evaluate the impact of vocabulary learning on comprehension.

Intervention types and contexts must also be considered when making generalizations for students with reading problems. There were many similarities among four of the five studies conducted with students with reading difficulties. Specifically, these studies used semantic mapping or semantic feature analysis, had moderate to high levels of discussion, and were conducted in small groups. Although many of the studies with general student populations included moderate to high levels of discussion, most were conducted at the classroom level, not small groups. Future research will need to be conducted to determine if other types of vocabulary instruction are beneficial to students with reading difficulties. Moreover, there may be an interaction between group size and discussion that is not fully captured by this set of studies and may need to be explored across different types of students. It should also be noted that although all of the studies with struggling readers showed similar effects, three of the five studies were conducted by the same research team (i.e., Anders, Bos, & Filip, 1984; Bos & Anders, 1990, 1992). The larger effects indicated for students with reading difficulties may be disproportionately influenced by something unique in how these researchers designed, implemented, and tested their interventions and may not be replicable with other research teams or across other populations of students with reading problems.

Future Considerations for Research and Practice

Although custom measures were sensitive enough to detect overall effects in comprehension and vocabulary, our ability to interpret this growth is restricted due to a lack of confidence in the reliability or validity of the measures used. Although the 2009 NAEP framework for the assessment of vocabulary and comprehension has yet to be tested, it offers a new perspective for the creation of conceptually sound and reliable measures. These guidelines, which include criteria for choosing words to assess and rules for creating items and distractors, may prove useful for experimenters designing custom as well as standardized measures.

The overall positive effect of vocabulary instruction on custom measures of comprehension highlights the importance of teaching vocabulary to promote understanding of text, especially for students with reading difficulties. Although no specific recommendations can be made for designing more effective vocabulary interventions to increase comprehension, results from the vocabulary measures suggest that practitioners should use high levels of discussion to promote vocabulary development. Researchers who want to better understand the factors related to vocabulary and its impact on comprehension need to systematically consider participant factors such as reading ability and grade level across intervention types, characteristics, and contexts. In addition, researchers

must use reliable and valid measures, fully report participant and intervention characteristics, and provide information concerning treatment fidelity. These efforts will put us closer to determining the optimal learning conditions under which vocabulary instruction is likely to impact students' comprehension.

ACKNOWLEDGMENTS

We would like to thank Mark Lipsey for his support and guidance on this project. This research was partially supported by grants from the US Department of Education (R305A070313, R305G050101, R305B040110).

REFERENCES

*References marked with an asterisk indicate studies included in the meta-analysis.

*Ahlfors, G. (1979). *Learning word meanings: A comparison of three instructional procedures.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

*Anders, P. L., Bos, C. S., & Filip, D. (1984). The effect of semantic feature analysis on the reading comprehension of learning-disabled students. In J. A. Niles & L. A. Harris (Eds.), *Changing perspectives on reading/language processing and instruction* (pp. 162–166). Rochester, NY: The National Reading Conference.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Apthorp, H. S. (2006). Effects of a supplemental vocabulary program in third-grade reading/language arts. *Journal of Educational Research, 100*, 67–79.

*Armstrong, J. C. (2000). *The integration of reading vocabulary techniques with scientific terminology in a sixth-grade classroom*. Unpublished doctoral dissertation, University of Connecticut, Storrs.

Askov, E. N., & Kamm, K. (1976). Context clues: Should we teach children to use a classification system in reading? *Journal of Educational Research, 69*, 341–344.

Barrett, M. T., & Graves, M. F. (1981). A vocabulary program for junior high school remedial readers. *Journal of Reading, 25*(2), 146–150.

*Baumann, J. F., Edwards, E. C., Font, G., Tereshinski, C. A., Kame'enui, E. J., & Olejnik, S. (2002). Teaching morphemic and contextual analysis to fifth-grade students. *Reading Research Quarterly, 37*(2), 150–178.

Baumann, J. F., Kame'enui, E. J., & Ash, G. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J. R. Squire, & J. Jensen (Eds.), *Handbook of reading research on teaching the English Language Arts* (2nd ed., pp. 752–785). Mahwah, NJ: Erlbaum.

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal, 107*(3), 251–270.

*Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*(4), 506–521.

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In A. Hiebert & M. Kamil, (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 223–242). Mahwah, NJ: Erlbaum.

*Blevins, M. L. (1970). *A comparative study of three methods of instructions in vocabulary on achievement of students in the Adult Institute.* Unpublished doctoral dissertation, Oklahoma State University, Stillwater.

*Block, C. C., & Mangieri, J. (2006). *The effects of Powerful Vocabulary for Reading Success on students' reading vocabulary and comprehension achievement.* New York: Scholastic Inc. Retrieved May 1, 2007, from http://teacher.scholastic.com/products/fluencyformula/ pdfs/ PowerfulVocab_Efficacy.pdf

Boettcher, J. V. (1983). Computer-based education: Classroom application and benefits for the learning-disabled student. *Annals of Dyslexia, 33*, 203–219.

*Bos, C. S., & Anders, P. L. (1990). Effects of interactive vocabulary instruction on the vocabulary learning and reading comprehension of junior-high learning disabled students. *Learning Disability Quarterly, 13*(1), 31–42.

*Bos, C. S., & Anders, P. L. (1992). Using interactive teaching and learning strategies to promote text comprehension and content learning for students with learning disabilities. *International Journal of Disability, Development and Education, 39*(3), 225–238.

Bos, C. S., Anders, P. L., Filip, D., & Jaffe, L. E. (1989). The effects of an interactive instructional strategy for enhancing reading comprehension and content area learning for students with learning disabilities. *Journal of Learning Disabilities, 22*(6), 384–390.

*Brabham, E. G., & Lynch-Brown, C. (2002). Effects of teachers' reading aloud styles on vocabulary acquisition and comprehension of students in the early elementary grades. *Journal of Educational Psychology, 94*(3), 465–473.

Bryant, P. D., Goodwin, M., Bryant, B. R., & Higgins, K. (2003). Vocabulary instruction for students with learning disabilities: A review of the research. *Learning Disability Quarterly, 26*, 117–129.

*Carney, J. J., Anderson, D., Blackburn, C., & Blessing, D. (1984). Preteaching vocabulary and the comprehension of social studies materials by elementary school children. *Social Education, 48*(3), 195–196.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs* (Vol. 129). Thousand Oaks, CA: Sage.

*Coyne, M., Zipoli, R., Jr., Loftus, S., & Kapp, S. (2007). *Direct vocabulary intervention in kindergarten: Investigating transfer effects*. Paper presented at the Institute of Education Sciences Research Conference, Washington, DC.

Davis, F. B. (1942). Two measures of reading ability. *Journal of Educational Psychology*, 33, 365–372.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*, 499–545.

*Dole, J. A., Sloan, C., & Trathen, W. (1995). Teaching vocabulary within the context of literature. *Journal of Reading*, *38*, 452–460.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170–177.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Freebody, P., & Anderson, R. C. (1983). Effects of text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior, 15*, 19–39.

Fukkink, R. G., & deGlopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, 68, 450–469.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics, 25*, 1–37.

Gipe, J. (1979). Investigating the techniques for teaching word meanings. *Reading Research Quarterly, 18*, 277–294.

Graesser, A., Golding, J. M., & Long, D. L. (1991). Narrative representation and comprehension. In D. Pearson, M. Kamil, R. Barr, & P. Rosenthal (Eds.), *Handbook of reading research* (Vol. 2, pp. 171–205). Hillsdale, NJ: Erlbaum.

*Hafner, L. E. (1965). A one-month experiment in teaching context aids in fifth grade. *Journal of Educational Research, 58*(10), 472–474.

Hayes, D. P., & Ahrens, M. (1988). Vocabulary simplification for children: A special case for "motherese"? *Journal of Child Language, 15*, 395–410.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499.

Hedges, L. V., & Olkin, D. (1985). *Statistical methods for meta-analyses.* San Diego, CA: Academic.

*Hogan, F. P. (1961). *Comparison of two methods of teaching word meaning through the use of word parts in grades ten, eleven and twelve.* Unpublished doctoral dissertation, Boston University, Boston.

*Hoisington, A. L. (1968). *An experimental investigation of a linguistic approach to vocabulary development which emphasizes structural analysis: Prefixes, suffixes and root words.* Unpublished doctoral dissertation, Washington State University, Spokane.

*Jackson, J. R., & Dizney, H. (1963). Intensive vocabulary training. *Journal of Developmental Reading, 6*, 221–229.

*Jones, S. T. (1984). *The effects of semantic mapping on vocabulary acquisition and reading comprehension of black inner city students.* Unpublished doctoral dissertation, The University of Wisconsin, Madison.

Jitendra, A. K., Edwards, L. L., Sacks, G., & Jacobson, L. A. (2004). What research says about vocabulary instruction for students with learning disabilities. *Exceptional Children, 70*, 299–323.

*Kameenui, E., Carnine, D., & Freschi, R. (1982). Effects of text construction and instructional procedures for teaching word meanings on comprehension and recall. *Reading Research Quarterly, 17*, 367–388.

*Kimer-Simon, K. L. (2003). *Storybook activities for improving language: Effects on language and literacy outcomes in Head Start preschool classrooms.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Klesius, J. P., & Searls, E. F. (1990). A meta-analysis of recent research in meaning vocabulary instruction. *Journal of Research and Development in Education, 23*, 226.

*Kux-Cox, J. A. (1974). *A comparison of two instructional methods utilizing the cloze procedure and a more traditional method for improving reading comprehension and vocabulary in context in a disadvantaged fourth-grade elementary school sample.* Unpublished doctoral dissertation, University of Southern Mississippi, Hattiesburg.

Lee, J., Grigg, W., & Donahue, P. (2007). *The Nation's Report Card: Reading, 2007* (NCES 2007–496). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

*Lieberman, J. E. (1965). *The effect of direct instruction in vocabulary concepts on reading achievement.* Unpublished doctoral dissertation, New York University, New York.

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

Lubliner, S. (2002). *The power of clarifying: A comparative analysis of strategies that strengthen comprehension.* Paper presented at annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Margosein, C. M., Pascarella, E. T., & Pflaum, S. W. (1982). The effects of instruction using semantic mapping on vocabulary and comprehension. *Journal of Early Adolescence, 2*(2), 185–194.

Marks, C. B., Doctorow, J., & Wittrock, M. C. (1974). Word frequency and reading comprehension. *Journal of Educational Research, 67*, 259–262.

Marmolejo, A. (1990). *The effects of vocabulary instruction with poor readers: A meta-analysis*. Unpublished doctoral dissertation. Columbia University Teachers College, New York.

Mastropieri, M. A., Scruggs, T. E., & Fulk, B. M. (1990). Teaching abstract vocabulary with the keyword method: Effects on recall and comprehension. *Journal of Learning Disabilities, 23*, 92–96.

McKeown, M. G., & Beck, I. L. (2006). Issues in the advancement of vocabulary instruction: Response to Stahl and Fairbank's meta-analysis. In K. A. Dougherty-Stahl & M. C. McKenna (Eds.), *Reading research at work: Foundations of effective practice* (pp. 262–271). New York: Guilford.

*McKeown, M. G., Beck, I. L., Omanson, R. C., & Perfetti, C. A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior, 15*(1), 3–18.

*McKeown, M. G., Beck, I. L., Omanson, R. C., & Pople, M. T. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20*, 522–535.

McNamara, D. S., & McDaniel, M. A. (2004). Suppressing irrelevant information: Knowledge activation or inhibition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*, 465–482.

*Medo, M., & Ryder, R. J. (1993). The effects of vocabulary instruction on reader's ability to make causal connections. *Reading Research and Instruction, 33*(2), 119–134.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53*, 253–279.

Nagy, W. (2005). Why vocabulary instruction needs to be long-term and comprehensive. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 27–44). Mahwah, NJ: Erlbaum.

*Nash, H., & Snowling, M. (2006). Teaching new words to children with poor existing vocabulary knowledge: A controlled evaluation of the definition and context methods. *International Journal of Language & Communication Disorders, 41*(3), 335–354.

National Institutes of Children's Health and Development. (2000). *Report of the national reading panel: Teaching students to read: An evidenced-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups.* Bethesda, MD: National Institutes of Health.

Otterman, I. M. (1955). The value of teaching prefixes and word-roots. *Journal of Educational Research, 48*, 611–616.

*Pace, R. E. (1986). *A comparison of two methods of teaching fifth grade science vocabulary: An imagery method and a traditional science textbook method.* Unpublished doctoral dissertation, University of Georgia, Athens.

Pany, D., & Jenkins, J. R. (1978). Learning word meanings: A comparison of instructional procedures. *Learning Disability Quarterly, 1*, 21–32.

*Pany, D., Jenkins, J. R., & Schreck, J. (1982). Vocabulary Instruction: Effects on Word Knowledge and Reading Comprehension. *Learning Disability Quarterly, 5*(3), 202–215.

Paris, S. G., & Stahl, S. A. (Eds.). (2005). *Children's reading comprehension and assessment*. Mahwah, NJ: Erlbaum.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly, 42*, 282–296.

Perfetti, C. (1985). *Reading ability.* New York: Oxford University Press.

Petty, W. T., Herold, C. P., & Stoll, E. (1968). *The state of knowledge about the teaching of vocabulary.* Champaign, IL: National Council of Teachers of English.

*Puhalla, E. M. (2005). *Teaching vocabulary from narrative and information text: Examining the effects of instructional intensity and judicious review on the vocabulary and expressive language performance of first-grade children at-risk of early reading difficulties.* Unpublished doctoral dissertation, Lehigh University, Bethlehem.

Raudenbush, S. W. (1994). Random effect models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–322). New York: Russell Sage Foundation.

*Reinking, D., & Rickman, S. S. (1990). The effects of computer-mediated texts on the vocabulary learning and comprehension of intermediate-grade readers. *Journal of Reading Behavior, 22*(4), 395–406.

*Roser, N., & Juel, C. (1982). Effects of vocabulary instruction on reading comprehension. In J. S. Niles & L. A. Harris (Eds.), *New inquiries into reading research. Thirty-first yearbook of the National Reading Conference*. (pp. 110–118). Albany, NY: National Reading Conference.

Ruetzel, D. R., & Hollingsworth, P. M. (1988). Highlighting key vocabulary: A generative-reciprocal procedure for teaching selected inference types. *Reading Research Quarterly, 23*, 358–378.

*Sampson, M. R., Valmont, W. J., & VanAllen, R. (1982). The effects of instructional cloze on the comprehension, vocabulary, and divergent production of third-grade students. *Reading Research Quarterly, 17*(3), 389–399.

*Schachter, S. W. (1978). *An investigation of the effects of vocabulary instruction and schemata orientation upon reading comprehension.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–284). New York: Russell Sage Foundation.

Shadish, W. R., Robinson, L., & Congxiao, L. (1999). *ES: A computer program for effect size calculation*. Memphis, TN: University of Memphis.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.

Spearitt, D. (1972). Identification of sub-skills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8*, 92–111.

*Stahl, S. (1983). Differential word knowledge and reading comprehension. *Journal of Reading Behavior, 14*, 33–47.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*, 72–110.

Stahl, S. A., & Nagy, W. (2006). *Teaching word meanings.* Mahwah, NJ: Erlbaum.

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

Stanovich, K. E., & Cunningham, A. E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *Journal of Educational Psychology, 85*, 211–229.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38*(8), 878–893.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*(10), 1046–1055.

*Thames, D. G. (1986). *Effects of prereading vocabulary strategies on vocabulary and comprehension of basal stories by primary children.* Unpublished doctoral dissertation, Louisiana State University, Baton Rouge.

*Thomas, C. L. (1998). *The effects of three levels of curricular modifications on the vocabulary knowledge and comprehension of regular education students and students with learning disabilities in content-area classrooms.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Tuinman, J. J., & Brady, M. E. (1974). How does vocabulary variance account for variance on reading comprehension tests? A preliminary instructional analysis. In P. Nacke (Ed.), *Twenty-third national reading conference yearbook* (pp. 176–184). Clemson, SC: The National Reading Conference.

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

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J., Butler, J., White, M. J., et al. (2005). Assessment of comprehension abilities in young children. In S. G. Paris

& S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 107–129). Mahwah, NJ: Erlbaum.

What Works Clearinghouse. (2007). *Technical details of WWC-conducted computations*. Retrieved August 21, 2007, from http://www.whatworks.ed.gov/reviewprocess/conducted_computations.pdf

Wittrock, M. C., Marks, C., & Doctorow, M. (1975). Reading as a generative process. *Journal of Educational Psychology, 67*, 484–489.

*Wixson, K. K. (1986). Vocabulary instruction and children's comprehension of basal stories. *Reading Research Quarterly, 21*(3), 317–329.

Wolfe, M. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 359–364.