Evidence-Based Reading Instruction for English Language Learners in Preschool through Sixth Grades: A Meta-Analysis of Group Design Studies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Insoon Han

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Kristen L. McMaster, Adviser

August, 2009

Acknowledgements

I was fortunate to study Special Education in the Educational Psychology Department at the University of Minnesota in which excellent professors and resources enabled me transit successfully from a student who had done merely well in school to a scholar who is ready to dedicate my life to educational research.

In particular, I am profoundly indebted to my adviser, Dr. McMaster, who was incredibly generous with her knowledge and time. She assisted me at every step, not just to complete the dissertation but to proceed throughout my whole graduate career, with her wealth of knowledge and wonderful personality. I am grateful to my dissertation committee members, Drs. Stanley Deno, Christine Espin, Lori Helman, and Geoffrey Maruyama, who served as midwives of my first piece of serious scholarship. Their insightful comments guided and enriched the meaning of my dissertation.

I acknowledge and thank the Graduate School for its generous offer of the Doctoral Dissertation Fellowship. The fellowship gave me an opportunity to complete my dissertation in a timely manner by devoting myself full-time to the effort of researching and writing of this dissertation.

Many thanks also go to Chin-Chih Chen, Viveca Braganza, and Seungsoo Yeo, all fellow doctoral students, and Heija Kang and Seunghee Kuk in Korea, for their friendship and encouragement.

Last, but not least, I want to add my thanks to authors of the studies included in my meta-analytic synthesis, who made it possible for my dissertation to be born.

Dedication

I dedicate my dissertation work to my family, especially…

to my husband, Duane Wamsley, for being always there for me; and

to my parents for their words of encouragement and push for tenacity.

Abstract

The purpose of the study was to synthesize research studies that examined reading instruction for English language learners (ELL) in preschool through sixth grades. Specific goals were to determine the availability of reading instruction for ELLs, the effects of that instruction, and which instructional programs can be considered evidence-based. An extensive search of the research studies identified 29 studies employing group design published in peer-reviewed journals since 1967, which yielded 44 samples with a total of 225 effect sizes. Findings revealed that at both tier 1 and 2, there are more than 10 instructional programs that address phonemic awareness and phonics instruction at preschool to 2nd grade levels, but there were no phonemic awareness and phonics instruction for third to sixth graders. The overall effect of reading instruction was moderate, with a mean ES of 0.50. HLM revealed that factors strengthen the effect were: English as a language of intervention; a session lasted 45 min or less; upper grade level; single grade; average or higher SES; and lower study quality. Three evidence-based or promising practices from 13 programs were identified; Keyword method, Proactive Reading and Peer-Assisted Learning Strategies. The research community should continue to reinforce efforts to improve the quality of research while seeking strong support for rigorous research from policy makers to empower ELL students who are otherwise likely to be at-risk readers by providing them with evidence-based reading instruction.

Table of Contents

List of Tables

CHAPTER I

INTRODUCTION

One of the significant questions in the current educational system in the United States (U.S.) is how to best educate children who come to schools with a home language other than English. For those young English language learners (ELLs), a primary role of the school is to help them learn to read and write proficiently in English. The challenges educators of ELLs currently face have been escalated by the following: the increasing number and diversity of ELL students, the significant literacy achievement gap between ELLs and their native English speaking (or English as first language; L1) peers, urgent calls for scientifically-based strategies or interventions in an era of educational reforms, the paucity of research available on the topic of English as second language acquisition for ELLs, and few research-based strategies and interventions for ELLs.

This chapter begins with a detailed discussion of these challenges. Then, the significance of the study is presented, with particular emphasis on the need for quantitative research synthesis on reading instruction for ELLs.

Problem Statement

Increasing Number of and Linguistic Diversity among ELLs

Reflecting the long history of immigration to the U.S. from hundreds of different countries, most U.S. classrooms have a substantial number of students of diverse ethnic, cultural, and linguistic backgrounds. For example, in 2006, among the 53.4 million students who were enrolled in elementary and secondary schools (5 to 17 years old), 10.9 million (20.4%) children spoke a language other than English at home. Among those 10.9 million ELLs, 2.8 million (25.7%) children spoke English less than "very well" (U.S. Census Bureau, 2005-2007 American Community Survey).

Whereas a number of these ELLs achieve English fluency each year, more and more new students lacking English fluency are entering school annually. The enrollment of ELLs has been growing rapidly in nearly every state across the country over recent years. For example, the percentage of ELLs enrolled in kindergarten through 12th grade nationwide between 1991 and 2002 has increased by 95%, compared to an increase of only 12% for the overall school-aged population (National Clearinghouse for English Language Acquisition [NCELA)], 2006). Twenty-five states saw the number of ELLs more than double from 1993 to 2003, and the number will only increase in the near future (DOE, 2009).

The linguistic diversity among the ELL population is also dramatically increasing. Spanish is by far the most common native language of ELLs, at approximately 76% (NCES, 2006); thus, it is not surprising that the majority of research on ELLs has focused on children from Hispanic backgrounds. However, ELLs are a heterogeneous population in terms of ethnicity, nationality,

socioeconomic background, immigration status, and generation in the U.S. (August & Hakuta, 1997) and the heterogeneity continues to increase. For example, in terms of linguistic diversity, five different Asian languages, Russian, and Arabic are in the top 10 languages, followed by Spanish. In Minnesota, with the large numbers of Southeast Asian immigrants, the Hmong language (34%) is as commonly spoken as Spanish (37%; Minnesota Department of Education, 2003).

The increasing linguistic diversity among ELLs becomes problematic for classroom teachers because the effects of first language on the development of English as a second language vary depending on what the first language is (e.g. Helman 2004). For example, empirical findings suggest that English decoding efficiency is determined, in part, by the first language-English orthographic distance (Koda, 2005). To illustrate, Spanish-speaking ELLs may take advantage of cognate relationships between Spanish and English when they learn to read in English, but ELLs with some other languages are not. Cognates are words that have similar spellings and meaning in two languages, such as "continue" in English and "continuar" in Spanish (Genesee, Geva, Dressler, & Kamil, 2006; Genesee, Lindholm-Leary, Saunders, & Christian, 2005). Cross-linguistic analyses also indicate that native Arabic and Japanese speakers are likely to encounter different types of word-level reading difficulties in English (Fender, 2003).

Reading Achievement Gaps between ELLs and L1 Students

A large number of ELLs struggle to learn to read or are underachieving in reading in English. According to the NCES, for example, in 2005, the percentage of 4th grade ELLs who performed below the basic reading level was 73%, compared to 33% of corresponding L1 students. In addition, as many as 93% of ELLs performed below the proficient reading level, compared to 66% of L1 students.

The gaps in reading test scores between ELLs and L1 students have been significant and persistent; for example, on the National Assessment of Educational Progress (NAEP) reading scale, ELLs in 4th grade, on average, scored 43 and 36 points lower than L1 students on a 0-500 scale in 1998 and 2007, respectively. The gaps among 8th graders are even larger: 46 and 42 points lower than L1 students in 1998 and 2007, respectively (NCES, 2008). Indeed, ELLs were among the farthest behind in this standardized reading test, with about 47% of 4th-grade ELLs and 51% of 8th graders behind their L1 peers. These data indicate that the reading achievement scores for one out of every two ELLs will have to improve to close these gaps (Fry, 2007). In fact, the gap between ELLs and L1 students is 3 to 18 points larger than the gaps between students who are and are not eligible for free or reduced-price lunch on the 2007 NAEP in reading and math.

The No Child Left Behind (NCLB) Act (2002) has raised public awareness of achievement gaps by legislating that *all* students must meet academic standards. "All" includes children who come from disadvantaged backgrounds, such as poverty, ethnic minority, disability, and limited English proficiency (LEP). That is, specific categories of students, including LEP students, must meet proficiency standards as a

group. Thus, the law seeks not only to raise the academic achievement of all students, but also to close the persistent achievement gaps between disadvantaged students and their peers. While LEP is the term used in state and federal laws, it can be viewed as a negative description of what students cannot do. English Language Learner (ELL) is a more positive term that has replaced LEP in many schools and in the literature. Despite this positive perspective, the prospect of closing the gap is not so positive for ELLs, given that the reading achievement gap between ELLs and L1 students has been significant and persistent for decades.

Urgent Call for Scientific, Research-Based Strategies and Interventions

NCLB requires schools to implement *scientifically-based* strategies that will benefit the majority of students and increase the likelihood that students will meet high academic standards. When schools fail to meet the standards, they are held accountable, and subject to consequences such as being required to provide their students with free tutoring or allowing students to transfer to a better-performing public school. The recent reauthorization of the Individuals with Disabilities Education Act (IDEA 2004) also calls for *scientifically-based* instruction. Further, IDEA allows local education agencies to use response-to-intervention (RTI) to identify students with learning disabilities (LD). RTI emerged as researchers, praticitioners, and policy-makers have seriously questioned the view that learning abilities and disabilities are inherent, rather than the result of a situational context. RTI may be particularly important for ELLs who are at risk for special education

referral and identification; indeed, empirical evidence suggests that, given proper instruction, some ELLs classified previously as LD can achieve grade-level norms (August & Shanahan, 2006). In other words, some ELLs identified incorrectly as LD are actually "instructionally disabled," meaning they have not received instruction appropriate for their unique language needs (Clay, 1987; Vellutino et al., 1996). This belief reflects one of the major rationales for an RTI approach (Fuchs & Fuchs, 2006).

According to the Council for Exceptional Children (CEC, 2007), the largest international professional organization in special education, the RTI process is designed to identify at-risk children early, to provide access to needed interventions, and to help identify children with disabilities. Specifically, RTI is a process intended to assist in identifying children with disabilities by providing data about how a child responds to scientifically-based intervention as part of the comprehensive evaluation required for identification of any disability, typically through three tiers of prevention. General education constitutes primary prevention, which is called as tier 1, in which high quality instructional programs are implemented. Students who fail to respond to this core instruction enter secondary prevention, or tier 2, by receiving supplementary, research-based small-group intervention. Students who respond poorly to this more intensive form of prevention become candidates for individualized tertiary intervention, or tier 3, and may undergo an evaluation to identify a specific disability. The goal of this approach is to ensure that quality instruction and remedial opportunities are available in general education, and that

special education is provided only for students who require more specialized services than can be provided in general education.

Determining the Evidence-Base of Interventions for ELLs

The above NCLB and IDEA mandates of implementing scientifically-based instruction raise concerns about instructional options for ELLs. Whereas much is already known about instructional principles that should be in place for all students, including ELLs, relatively little is known about scientific, research-based interventions for this specific subgroup. For example, the Center for Research on Education, Diversity and Excellence (2004) published a set of successful teaching strategies for all students, including ELLs, and Bear et al. (2007) organized research findings on effective instructional strategies for ELLs. According to their findings, as a whole, instruction for ELLs is most effective when it is explicit and systematic, when it allows ELLs to interact with others in their learning community, when it helps the students make connections to what they already know, and when they actively construct knowledge (Bear et al.) Thus, teachers can integrate these general principles into their daily instruction to support ELL students. Moreover, instructional programs or interventions that have incorporated some or all of these principles for ELLs have been researched, published, and reviewed. However, because of the mismatch between how these studies have been synthesized in existing reviews and the definition of evidence-based practice in education, few interventions for ELLs that qualify as “evidence-based” have been identified.

An important question, then, is this: what does it means to call an intervention *scientifically-based* (which is the term used in legislation) *or evidence-based* (which is the term commonly used in the field of education)? Although there has been a great deal of debate on this issue, standards for evidence-based practice have been proposed in general education (the What Works Clearinghouse [WWC], 2002), in school psychology (Task Force on Evidence-Based Interventions in School Psychology, 2003), and in special education (e.g., Gersten et al., 2005; Horner, Carr, Halle, & Odom, 2005). Although refining some criteria is still under discussion, these standards are generally accepted by the education community (e.g., Cook, Tankersley, & Landrum, 2009).

In this study, I used the criteria proposed by Gersten et al. (2005) for identifying evidence-based practice from studies employing group experimental designs, because this criteria is the latest one and has been developed by examining the strengths of and criticisms for previous standards to enhance the credibility. These criteria are: "at least two high quality studies, or four acceptable quality studies that support the practice; *and* the weighted effect size greater than zero" (Gersten et al., p. 162). By these criteria, to be considered an evidence-based practice, the practice should be supported by a certain number of studies with high or acceptable quality and study outcomes should be aggregated to calculate a weighted effect size. To yield a weighted effect size across studies, a meta-analysis should be conducted as a method of synthesizing the literature. The reason for this is that meta-analysis can provide the average effect of a given intervention weighted by the

studies. Furthermore, it can examine quantitatively the pattern of the effect, such as which intervention works with whom, by whom, and in what context.

To date, researchers have conducted two syntheses on reading instruction for ELLs (Cheung & Slavin, 2005; Klingner, Artiles, & Barletta, 2006)[1]. The methodology and findings of these two research syntheses will be discussed in greater detail in Chapter 2; however, it is important to note that the synthesis conducted by Cheung & Slavin provided a weighted effect size, but did not report on study quality. Because Klinger et al.'s review was narrative, neither effect size nor study quality was reported. Thus, neither study could provide conclusive answers about reading instruction for ELLs. A meta-analysis combined with the examination of study quality was used in the present research as a method of research synthesis to examine and identify evidence-based practices.

Significance of the Study

Methods of Research Synthesis

A single study hardly ever provides a definitive answer to a research question in the field of education (Hunter, Schmidt & Jackson, 1982; Suri, 2000). Results of a single study are often influenced by sampling characteristics such as the

---

[1] A third synthesis has been conducted by the WWC, but they did not release it until I had completed analyses for my dissertation. Because of its similarity to my review, I included a special section in the Discussion (Chapter V) to compare and contrast results of the WWC and my own meta-analyses.

sample size and study setting. Consequently, a comprehensive investigation requires the combination of results from several individual studies. As research results accumulate, however, different individual studies often provide conflicting results, which increases the difficulty of interpretation of cumulative results (Rudner, Glass, Evartt, & Emery, 2002). Therefore, synthesis of research across a large body of empirical studies is required to accumulate knowledge with certainty.

The major roles of research synthesis are to transmit the accumulated knowledge and determine the direction of further research, practice, policies. Given that both transmission of knowledge and providing direction for the field are critical in education, the methodology of research synthesis is crucial (Glass, McGaw & Smith, 1981). The most frequently used methods of synthesizing research in education are traditional narrative review, meta-analysis, and best-evidence synthesis. Each methodology has its own strengths and weaknesses (Suri, 2000, Suri & Clarke, 2009).

Narrative reviews. Traditional narrative reviews of research are flexible in their methodology, including the inclusion criteria and the decision rule of effect. Although an experienced research reviewer can synthesize studies effectively by adopting this method, the flexibility in the methodology can result in misleading conclusions mainly because of its high level of subjectivity, potential for false negative, and no weight for sample size and effect size. Moreover, these methods often fail to identify study characteristics that could moderate the effect (McGaw, 1997, Suri, 2000; Suri & Clarke 2009).

Meta-analysis. Meta-analysis, a quantitative method of research synthesis proposed by Glass (1976), has several advantages over traditional narrative reviews. It includes all quantitative empirical studies on a given topic, so should be comprehensive and free from subjectivity introduced by selective inclusion of studies. It also offers more conclusive answers than narrative reviews by providing the direction and magnitude of effects across studies, and more sophisticated answers by identifying moderator variables (Suri, 2000; Lipsey & Wilson, 2001).

Yet, meta-analysis has been criticized for overgeneralization, which can result from including poorly designed studies and consequently ignoring the possible impact of study quality (also referred to as the "mixing apples and oranges" problem), and selection bias such as the exclusion of qualitative studies (for which effect size cannot be computed (Eysenck, 1994). The strengths and weaknesses of meta-analysis will be discussed in greater detail shortly.

Best-evidence synthesis. Slavin (1986, 1987) proposed the method of "best-evidence synthesis" as an alternative to traditional narrative and meta-analytic reviews. By using well-specified a priori inclusion criteria, the best-evidence synthesis includes only the "good or best-evidence" studies. Like meta-analysis, this method uses the quantification of effect sizes and systematic inclusion procedures. Moreover, like narrative review, this method allows flexibility by giving attention to individual studies and methodological issues. Thus, the statistical analysis of quantitative findings is supplemented with a rich literature review that explains any discrepancies observed that cannot be quantified.

However, some researchers have questioned the idea that meta-analyses should use only "good or best-evidence" studies (e.g., Glass, 2000). They argue that meta-analyses must deal with all studies, good and bad, and that studies' results are only properly understood in the context of each other, not after having been censored by some *a priori* set of inclusion criteria or "prejudices" under the name of "best-evidence" (Glass). The best-evidence synthesis has also been criticized in that its results may summarize only a narrow research domain such as studies employing randomized designs, resulting in limited generality to other research domains--for example, quasi-experimental studies. Another option is to include both research domains in a meta-analytic review, and to examine the design feature (i.e., experimental vs. quasi-experimental) as a factor affecting the strength of effectiveness (Lipsey & Wilson, 2001).

In the context of the criteria for evidence-based practices in special education proposed by Gersten et al. (2005; i.e., a certain number of quality studies and weighted effect size greater than zero), meta-analysis is most appropriate if it is supplemented with the examination of study quality. A traditional narrative review does not yield a weighted effect size. Although best-evidence synthesis produces weighted effect sizes, this method cannot identify evidence-based practice; instead, best-evidence is a priori inclusion criteria defined by a researcher. By calculating effect size *and* examining quality of studies included, meta-analysis can be used to identify evidence-based practices.

Meta-Analysis as a Method of Research Synthesis

The definition of meta-analysis (Glass, 1976) is “the statistical analysis of a large collection of analysis results for the purpose of integrating the findings” . Thus, meta-analysis is the statistical method that is used in a meta-analytic review. The basic purpose of meta-analysis is to provide the same methodological rigor to a literature review as we require from experimental research (DeCoaster, 2004). Since the psychologist Gene Glass coined the term ‘meta-analysis’ in a paper entitled ‘Primary, Secondary and Meta-analysis of Research’ in 1976 , it became a popular method in agricultural sciences, medicine, psychology, and education mainly because of its useful system of rules that guide the process of summing up research findings (Banda & Therrien, 2008). There has been a growing consensus that meta-analysis is a useful and important tool for synthesizing research literature in educational research (Lipsey & Wilson, 2001; Harwell & Maeda, 2008; Hunter & Schmidt, 2004).

The strengths of meta-analysis are four-fold: meta-analysis (1) represents findings in a more differentiated and sophisticated manner than conventional reviews, (2) is capable of finding relationships across studies that are obscured in other approaches, (3) protects against over-interpreting differences across studies, and (4) can handle a large number of studies, which might be overwhelming in traditional reviews (Banda & Therrien, 2008; Lipsey & Wilson, 2001).

In addition to requiring a good deal of effort, meta-analysis has several weaknesses or has been criticized for the following reasons: (1) mechanical aspects

do not lend themselves to capturing more qualitative distinctions among studies; (2) meta-analysis "mixes apples and oranges," which means that by aggregating incommensurable study findings such as including poorly designed studies, the mean effect size and other summary statistics produced by meta-analysis are not meaningful; (3) most meta-analyses include studies that are flawed to one degree or another; for example, equivalence between groups in a randomized design can be destroyed because of severe and/or differential attrition; and (4) publication bias poses a threat. For example, because negative and null finding studies are often not reported or published; thus, meta-analyses that include only published studies are quite likely to introduce an upward bias to the average effect size (Banda & Therrien, 2008; Lipsey & Wilson, 2001).

Evidence-Based Practice and Study Quality

In the era of NCLB and the reauthorized IDEA, classroom teachers must employ scientific, research-based strategies and interventions for their students, including ELLs. Evidence-based practices in general and scientifically-based intervention in particular are hot topics and top priorities in educational policy and research. A large space allowance concerning evidence-based practices in major educational journals reflect the importance of this topic in both general and special education (see special issues in *Educational Researcher, 37*, 2008; *Exceptional Children, 75*, 2009).

As mentioned earlier, Gersten et al. (2005) proposed the criteria for concluding a practice as *evidence based*, in special education. The criteria are: "(1) There are at least four acceptable quality studies, or two high quality studies that support the practice; *and* (2) the weighted effect size is significantly greater than zero" (p. 162). They further specify essential and desirable indicators of study quality.

To be considered acceptable quality, a study must meet all but one of the e*ssential* quality indicators and at least one of the *desirable* quality indicators specified by Gersten et al. (2005; see Appendix A). To be considered high quality, a study must meet all but one of the essential and at least four of the desirable indicators. The reason Gersten et al. stressed the weighted effect size was that this statistic takes into account the number of studies conducted on the specific intervention, the number of participants in the studies, and the magnitude and consistency of effects. They also proposed criteria for considering a practice as *promising:* (1) same as the first criteria for evidence-based *and* (2) there is a 20% confidence interval for the weighted effect size that is greater than zero.

For the above criteria, Gersten et al. (2005) weighed heavily the quality of the research design, based on findings of Simmerman and Swanson (2001). Simmerman and Swanson documented the impact of specific flaws in research design on a study's effect size, and argued that the higher quality a study has, the lower its effect size. In other words, better controlled studies, such as those using measures of documented reliability and validity, controlling for effects of teachers or

interventionists, and using the appropriate unit of analysis in statistical calculations, appear to be less biased in favor of the intervention, which results in lower effect size. Simmerman and Swanson stated that the emphasis on study quality goes hand-in-hand with the current emphasis on scientifically based research.

Purpose

The purpose of this study was to identify evidence-based practices in reading instruction for ELLs. To achieve this goal, the examination of study quality was treated as equally important as effect size, because, according to Gersten et al. (2005), both are necessary criteria for concluding that a practice is evidence-based. Furthermore, by including a close examination of study quality, the present meta-analytic review is expected to substantially address most of the weaknesses of meta-analysis mentioned above. Specifically, the examination of study quality along with effect size statistics allows this review to make more qualitative distinctions among studies; investigate study design and quality as a factor affecting the strength of effectiveness; quantify the rigor of design, including presence of differential attrition; and examine empirically the relation between study quality and effect size.

Definition of Key Terms

The following key terms are provided in order to maintain clarity in describing certain phenomena related to the reading instruction for ELLs.

English language learners (ELLs): ELLs are students with a primary language other than English who have a limited range of reading in English. ELLs also include students identified and determined by their school as having limited English proficiency (LEP) and a language other than English spoken in the home (U.S. Department of Education, Institute of Education Sciences, 2009).

Key Instructional Components of Reading: The National Reading Panel (NRP, 2000) concluded that there are five key components to an effective reading program. These components include; phonemic awareness (PA), phonics (p), fluency (f), vocabulary (v), and comprehension (c).

Reading Outcome Measures: Reading outcome measure were defined as quantitative measures of English reading performance, such as standardized tests and informal reading inventories, on any of the five components defined above.

Meta-analysis: Statistical analysis of a large collection of findings from several studies for the objective of integrating the findings (Glass et al., 1981).

Effect size: A statistic that encodes the critical quantitative information from each relevant study findings (Lipsey $ Wilson, 2001).

CHAPTER II

LITERATURE REVIEW

The National Reading Panel (NRP, the National Institute of Child Health and Human Development [NICHD], 2000) identified phonemic awareness, phonics, fluency, vocabulary, and text comprehension as the five key components of reading. By recognizing the importance of teaching language-minority students to read and write proficiently in English, the U.S. Department of Education's Institute of Education Sciences (IES) created the National Literacy Panel on Language-Minority Children and Youth. The formal charge of the panel was to identify, assess, and synthesize research on the education of language-minority children and youth regarding to literacy attainment and to produce a comprehensive report (August & Shanahan, 2006). One of their major findings was that instruction providing substantial coverage in the key components of reading identified by the National Reading Panel has a positive influence on the literacy development of ELL students, just as for L1 students.

Whereas most instructional practices that are effective for L1 populations were effective for ELLs, the National Literacy Panel (August & Shanahan, 2006) suggested that adjustments to these approaches are needed to maximize benefits for ELLs. For example, young Spanish-speaking ELL students might make the best progress when they are taught particular phonemes and combinations of phonemes

in English that do not exist in Spanish. Moreover, the panel reported that some instructional approaches found to be successful with L1 students may not have as positive of a learning impact on ELLs. For example, it does not appear to be sufficient to teach ELLs reading skills alone; rather extensive oral English development must be incorporated into literacy instruction. Thus, educators cannot guarantee that all instruction designed with and/or for L1 students will be equally effective for ELLs, due to their special language learning needs related to the process of second-language acquisition.

In this chapter, I begin by reviewing theories of second language acquisition to provide background understanding of ELLs' reading development in English. Then, I review previous efforts to synthesize research on reading programs for ELLs.

Theories of Second Language Acquisition and ELLs' Reading Development

When ELLs experience difficulties learning to read, a question commonly raised is this (Case & Taylor, 2005; Klingner et al., 2006): Do they struggle to learn to read due to second language acquisition or to specific learning disabilities (LD)? To answer this question, educators need to understand both the natural process of second language acquisition and the characteristics of LD. Yet, ELLs who are in the natural process of second language development and students with LD share some common characteristics. For example, in the case of reading, they may make similar types of errors in pronunciation, syntax, and semantics (Case & Taylor). Therefore,

it is difficult for teachers with ELLs who are struggling to read to differentiate between the process of second language acquisition and the characteristics of LD.

Due to this difficulty, on the one hand, many teachers might hesitate to refer ELLs at-risk of reading failure to special education (Klingner et al, 2006). As a consequence, some ELLs with actual LD are inappropriately assumed to have only second language acquisition-related difficulties, resulting in under-identification of ELLs who need special education. On the other hand, some ELLs who have reading difficulties due to second language acquisition are assumed incorrectly to have LD, which results in over-identification of ELLs as needing special education (McCardle, Mele-McCarthy, Cutting, Leos, & D'Emilio, 2005). Thus, prior to referring an ELL student to special education, teachers need to understand key concepts of second language acquisition.

An understanding of second language acquisition may not only foster proper referral to special education, but also improve the ability of general classroom teachers to teach ELLs in their classrooms (Fillmore & Snow, 2002; McCardle, Mele-McCarthy, & Leos, 2005). Theories of second language acquisition in general and reading development in English in specific are helpful in promoting such an understanding. In particular, understanding these theories can help teachers develop appropriate instructional strategies that guide students along a continuum of language development (Northwest Regional Educational Laboratory, 2003). The most prominent of these theories are reviewed below.

Cummins's Theory of Second Language Acquisition

Cummins (1981) distinguished between two types of language: basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP). BICS involve the informal language of conversation, often referred to as the "language of the playground" in that most children learn BICS through informal interaction with their peers. Cummins suggested that the acquisition of this level of communication takes 2 to 3 years of exposure. However, developing fluency in more technical, academic language—CALP--can take a much longer time. CALP refers to language skills that are associated with literacy and cognitive development. As opposed to BICS, these skills are learned most often through formal instruction. Because children generally gain CALP while at school, it takes much longer to develop. To obtain CALP, it takes an average L1 learner 4 to 7 years (Cummins) and bilingual children 4 to 9 years (Collier, 1989).

Krashen's Theory of Second-Language Acquisition

Krashen's (1987, 1988) theory of second-language acquisition consists of five hypotheses: acquisition-learning hypothesis, monitor hypothesis, natural order hypothesis, input hypothesis, and affective filter hypothesis. Among these, the most relevant to the current review is the natural order hypothesis.

According to the natural order hypothesis, the order of acquisition of language skills is predictable and on a continuum of five sequential stages: the silent/receptive or preproduction stage, the early production stage, the speech

emergence stage, the intermediate language proficiency stage, and the advanced language proficiency stage. Because the characteristics of these stages correspond to those of ELLs' stages of English reading development in terms of concept and sequential order, detailed explanation of each stage will be presented in the section on ELLs' stage of reading development below. When teachers understand that students go through a predictable and sequential series of second language developmental stages, they can adequately predict and accept a student's current stage and modify their instruction to encourage progression to the next stage (Reed & Railsback, 2003).

The Stages of ELLs' Reading Development in English

Researchers (Chall, 1983; Ehri, 1997) have described the progression of reading stages in English (Bear, Invernizzi, Templeton, & Johnston, 2004), which correspond well with the stages of second language development in terms of concepts and sequential order. These stages overlap because learning is gradual and learners do not abruptly move from one stage to the next. Individuals also may vary in their rate of progression through these stages, but most tend to follow the same order of development: the emergent stage, the beginning stage, the transitional stage, the intermediate stage, and the advanced stage (Bear et al., 2004; Bear, Helman, Templeton, Invernizzi, & Johnston. 2007).

The emergent stage for ELLs typically corresponds to ages 1 to 7 years (pre-K to mid-first grade) or is evident in older children without formal schooling. The

child undertakes reading in earnest, but adults recognize their efforts as more *pretend* than real. For example, children may call out the name of a favorite fast food restaurant when they recognize its logo, but lack an understanding of the alphabetic principle. Thus, reading instruction in this stage should develop phonemic awareness and alphabetic knowledge with pictures, songs, games, and matching activities. Explicit instruction is required to build English vocabulary (Bear et al., 2004; 2007).

The beginning stage corresponds to age 4 to 9 years (kindergarten to early third grade) or is evident in older students with limited formal schooling. Beginning readers have moved from pretend reading to real reading. The understanding of the alphabetic nature letter-sound relation, of English is a major challenge for these readers. Therefore, they require much support in the form of predictable texts and much teacher guidance in the form of strategies to use for decoding. They benefit from repeated reading of predictable texts with many phonetically regular words and high frequency words, but need explicit phonics instruction for sounds that are difficult in English. Again, building vocabulary should be a part of early reading instruction at this stage (Bear et al., 2004; 2007).

The transitional stage corresponds to ages 6 to 12 years (first to mid-fourth grade) or is evident in older students who began literacy instruction in English at a later age. Student in this stage still need support for long vowels. However, their reading begins moving from word-by-word to phrase-by-phrase with greater expression, so they approach fluent reading at their instructional level. Content-area vocabulary instruction embedded in language-rich activities is needed, and cognates

in students' home language can be incorporated into vocabulary instruction (Bear, 1992; Bear et al., 2004; 2007).

The intermediate stage corresponds to ages 8 to 18 years, (third to eighth grade) or is evident in students at later grades who came more recently to English reading instruction. Intermediate readers read most texts with accuracy and speed, both orally and silently. Thus, they begin reading to learn (such as content area material). Reading instruction should investigate academic vocabulary in English. Instruction continues to use cognates in students' home language to increase vocabulary. For these students, success in reading is related to familiarity and experience with the topic being discussed; thus, reading materials related to their ethnic and cultural background can be used for fostering comprehension (Bear et al., 2004; 2007).

The last stage, the advanced stage, corresponds to ages 10 years or older, or fifth to twelfth grade. Advanced readers have broad experiences that allow them to choose among a variety of reading materials to suit their purpose for reading and their own interests. Their vocabulary and comprehension increase with reading. Again, instruction can use cognates in students' home language to increase vocabulary (Bear et al., 2004; 2007).

Summary

ELL students go through a predictable and sequential series of English language developmental stages, which correspond to second language acquisition

stages in terms of concepts and sequential order. Thus, it is important for educators to understand these stages. This understanding helps educators predict and accept an ELL's current stage of learning and develop and modify their instruction. When a ELL student has reading difficulties, his/her teacher might consider evaluating the student's ELL language development stage first, rather than basing expectations on his/her grade or age. In short, ELLs' instructional needs vary, and Teacher of ELLs are strongly encouraged to use literacy assessments to identify students that require additional support, and provide enriched learning opportunities at their developmental level (Helman 2005). For example, if a fifth-grade age-level Hmong student immigrated to the U.S. recently and never received formal schooling in his home country, his teacher might expect him to be at the emergent stage of language and reading acquisition and provide instruction accordingly. If this instruction is not successful in helping the student improve English and reading skills, moving to more intensive intervention should be considered.

Previous Research Syntheses on Reading Programs for ELLs

To date, there have been several research efforts to synthesize the effects of instruction for ELLs. These research syntheses fall into three broad categories: (1) narrative reviews, which include reviews by Baker and de Kanter (1981), Rossell and Baker (1996), Genesee et al. (2005), and Klingner et al (2006); (2) meta-analyses, which include reviews by Greene (1997), Willig (1985), and Rolstad,

Mahoney, and Glass (2005); and (3) best-evidence syntheses, which include reviews by Slavin and Cheung (2005), and Cheung and Slavin (2005). All but two of these research syntheses have focused on general educational outcomes of bilingual programs for ELLs. Only Slavin and Cheung (2005), Cheung and Slavin (2005) and Klingner et al. (2006) conducted empirical literature reviews focusing specifically on reading instruction for ELLs; thus, two reviews are described in detail below (because studies included in the two reviews by Cheung and Slavin were the same, I will discuss only one (Cheung & Slavin, 2005) below.

Cheung and Slavin's (2005) Synthesis

Cheung and Slavin (2005) conducted a best-evidence synthesis of experimental studies comparing bilingual and English-only reading programs for ELLs in elementary and secondary grades. The dependent variables were quantitative measures of English reading performance, such as standardized tests and informal reading inventories. Among a total of 17 studies that met inclusion criteria, 13 focused on elementary reading for Spanish-dominant students. Of those, 9 favored bilingual approaches on English reading measures and 4 revealed no differences. The weighted effect size was 0.33, which was statistically significantly different from zero. The definition of quality of study in this review was the design and length of conducting the study; randomized multiyear studies were listed as the highest quality, followed by matched multiyear studies; matched 1-year studies were listed as the lowest quality. Among the 13 studies on elementary reading, 3 studies

were listed as highest quality, 8 moderate, and 2 as the lowest quality. The authors concluded that although the number of high-quality studies was small, the evidence of the review favored bilingual approaches, especially paired bilingual strategies that involve teaching reading in the students' native languages and English at different times each day.

Although this review contributed substantially to our understanding of the availability and overall effectiveness of reading instruction for ELLs in elementary grades, it did not produce any information about the impact of contextual factors on the effectiveness of that instruction. Moreover, Cheung and Slavin's definition of study quality as the design and length of study is somewhat limited. Although design and length can be considered as proxies of study quality, the definition is much less comprehensive than the one recently proposed in special education (Gersten et al., 2005); in short, with its definition of study quality, the researchers did not comprehensively examine and identify evidence based-practices in reading for ELLs.

Klingner, Artiles, & Barletta's (2006) Synthesis

Klingner et al. (2006) reviewed 23 studies to differentiate between ELLs who struggle because of language acquisition and those who have LD, but only seven studies were categorized as providing data on interventions; of these, five studies investigated intensive reading interventions with ELLs who showed early signs of struggling, and two were on reading comprehension strategy instruction. Among these, early intervention programs that combined phonological awareness

and other reading activities with ESL strategies were considered the most promising.

This review contributed substantially to our understanding of which interventions work for ELLs in what context, by categorizing studies in terms of population characteristics and subtype, the role of context in understanding ELLs' struggles, referral issues, assessment practices, and the predictors of reading achievement. Despite the rich descriptive information provided in this review, it included only seven studies that provided data on interventions. The number of studies included was insufficient to conduct meta-analysis; indeed, the review was narrative, so neither effect sizes nor information about study quality was reported. Related to this limitation, the authors of this review stated that further research is needed before firm conclusions can be drawn; in particular, the authors stated that a better understanding of the role of native language instruction is needed, and additional research is needed to know about what specific approaches work best with whom and under what circumstances.. A meta-analytic review that includes identification of moderators of instructional effects would be an appropriate approach to examine these unsolved questions.

What We Still Need to Know about Evidence-Based Reading Instruction for ELLs

The research syntheses described above have provided a substantial amount of information about the availability of reading instruction for ELLs, especially those in elementary grades. Still, many important questions remain; among these are (1) Which reading instructional practices can be considered "evidence-based" in light of

recent criteria established for educational research? (2) What contextual factors impact the effectiveness of instruction for ELLs? and (3) What is the relation between study quality and effect size? First, the value of identifying evidence-based instruction cannot be overstated in an era of standards and accountability as mandated by NCLB. Second, even evidence-based instruction may not necessarily work for all students in all educational settings. Interactions between study characteristics, intervention effects, and contextual factors are crucial information for teachers who seek effective instruction for their classrooms, which are situated in diverse educational environments. Finally, the examination of study quality will enable researchers to empirically approach the "mixing apples and oranges problem" by determining the relation of study quality to effect size.

Study Purpose

The purpose of this study was to conduct a meta-analytic review to examine the *availability* of reading instructional practices, and the *effectiveness* of those available practices, and to identify *evidence-based instruction* for ELLs in elementary grades. The first goal was to examine the availability of instruction in the context of current educational reforms (i.e., NCLB's [2002] focus on reading and RTI as allowed by IDEA[2004] ; thus, availability is examined in terms of (a) what instruction is available for each tier of the RTI framework; (b) which key instructional components of reading were addressed, including phonemic awareness,

phonics, fluency, vocabulary, and text comprehension (National Reading Panel, 2000); and (c) other major study characteristics, such as intervention descriptors (e.g., intervention language), and participants descriptors (e.g., grades).

The second goal was to examine the effectiveness of instruction in terms of the following aspects: (a) overall effectiveness; (b) moderators affecting the strength of study effects; and (c) effectiveness broken down by significant moderators.

The third and last goal was to identify evidence-based reading instruction using the following steps: (a) evaluate the quality of the studies by using the quality indicators established by Gersten et al. (2005); (b) identify scientifically-based reading instruction based on the two criteria of evidence-based practice proposed by Gersten et al.: the quality of study and weighted effect size; and (c) examine the relation between these two criteria.

CHAPTER III

METHODOLOGY

The purpose of this study was to conduct a meta-analytic review on reading instruction for ELLs in elementary grades. Once research problems have been specified, the general steps in conducting a meta-analytic study are: (1) searching the literature, (2) selecting studies, (3) coding study characteristics, (4) computing and coding effect sizes, and (5) conducting data analyses. This chapter describes all 5 steps in order.

Searching the Literature

To locate all of the existing research on reading instruction for ELLs, I used three modes of searching: (a) primary searches in major indexes in the area of education and psychology, such as ERIC, Education Full Texts, and PsycINFO by using three categories of key words: ELL, LEP, and language minority for participants; reading, literacy, and language for content area; and instruction, intervention, and strategies for instruction, (b) secondary searches by tracing back to citations and browsing major journals, including *Bilingual Research Journal, Exceptional Children, Education and Treatment of Children, Journal of educational Psychology, The Elementary School Journal, Journal of Learning Disabilities,*

Learning Disability Research & Practice, Learning Disability Quarterly, Journal of Literacy Research, Reading Research Quarterly, Reading Teachers, Remedial and Special Education, Review of Educational Psychology, School Psychology Review, Journal of Special Education, and *Scientific Studies of Reading*, and (c) ancestral searches of studies identified from reference lists of previous literature reviews and obtained studies.

Selecting Studies

To determine which studies to include in this meta-analysis, the following six criteria were used. The studies must have:

(1) included English language learners (ELLs) as participants. By ELLs, I mean students who speak another language than English, who are in the process of acquiring English as a second or additional language, and who have not yet achieved full English proficiency;

(2) focused on reading instruction in which reading outcome measures were included. By definition, reading instruction includes instructional activities in English and/or ELL students' home language on at least one of five key reading components, which are phonemic awareness, phonics, fluency, vocabulary, or text comprehension. Outcome measures were defined as quantitative measures of English reading performance, such as standardized tests and informal reading inventories, on any of the five components. Examples of instructional activities and outcome

measures are presented in Appendix B.

(3) were conducted with a Pre-Kindergarten to sixth-grade population. Although most research syntheses on instruction include studies targeted on grades K-6 (e.g., Yang, 2006), the present review included studies on preschool children because the theories on second language acquisition address the importance of instructional attention to the emergent stage, which corresponds to children 1 to 7 years old, or preschool to first-grade level, with emphasis on phonemic awareness and phonics instruction.

(4) used group experimental or quasi-experimental designs. Studies employing single-subject designs were not included because single-subject effect size statistics are different from effect size statistics of studies with group designs. Although multiple meta-analyses combining studies from different designs can be performed on the same body of literature (Lipsey & Wilson, 2001), this approach has been rarely attempted in the literature mainly because very few studies provide sufficient information to calculate effect sizes across studies of different designs (DeCoster, 2004).

(5) were conducted in English-speaking countries, in which English is the primary language of instruction in mainstream schools.

(6) published in peer-reviewed journals since 1967. The 1967 cutoff was chosen with the intention of including studies over the last 40 years.

Coding Study Characteristics

Each study was be coded by research design and quality descriptors, study descriptors, intervention descriptors, and participant descriptors, using a detailed coding scheme and coding manual (Appendix C). The coding scheme was as follows:

Research design and quality descriptors: (a) research design (experimental and quasi-experimental), (b) comparison groups (ELL treatment vs. ELL control, ELL+L1 vs. ELL+L1, ELL vs. L1, ELL at-risk vs. ELL not-at-risk) (c) the quality indicators (see Appendix A; present or not present).

Intervention descriptors: (a) title of instructional practice or intervention, (b) language of intervention (English, Spanish, other), (c) key instructional components (phonemic awareness, phonics, fluency, vocabulary, comprehension, other), (d) number of key components included, (e) preventive level of tiers in RTI framework (primary, secondary, tertiary) (f) grouping (class-wide, small group, pair, one-to-one), (g) length of each session (min), (h) frequency per week, (i) a length of intervention (months), (h) interventionist, (i) training for interventionist, (j) place, (k) assistance or consultation, (l) fidelity, and (m) interrater agreement for fidelity evaluation.

Participant descriptors: (a) grade, (b) gender, (c) race/ethnicity, (d) home language, (e) SES, (f) English language proficiency level, and (g) reading ability level (high, average, low or below grade or at-risk, learning disabilities, other).

Outcome measure descriptors: (a) construct of reading performance (phonemic awareness, phonics, fluency, vocabulary, comprehension, other), (b)

standardization (standardized or raw scores), (c) alignment with instructional components (match or mismatch), and (d) timing (immediately after the end of intervention or follow-up).

Interrater Agreement

To establish interrater agreement (IRA), the author served as the first rater and a recent graduate of the University of Minnesota Special Education doctoral program served as a second rater. First, the two raters coded two studies together, using the coding manuals for training purposes. Codes with low IRR were identified, discussed and rewritten in greater detail. The two raters then independently coded five studies selected randomly from the 29 studies. IRA was used for coding both study characteristics and quality indicators. IRA was defined as the frequency of the occurrence of agreement out of all possible opportunities to agree. The overall IRA for the study characteristics was 84%; specifically, IRA was 100% for participant descriptors, 80% for intervention descriptors, and 72% for outcome descriptors. The overall IRA for quality indicators was 86%.

For coding of quality indicators, the Kappa coefficient proposed by Cohen (1988) for qualitative or categorical items was also used. The Kappa was not used for the above study characteristics as some variables were on a continuous scale. The value of Kappa was defined as (observed agreement - expected agreement) / (100% - expected agreement). Kappa ranges from -1.0 to 1.0; larger numbers mean better reliability. The results of the IRA analysis using SPSS was Kappa = 0.61 (p = 0.00).

As a rule of thumb, values of Kappa from 0.40 to 0.59 are considered moderate, 0.60 to 0.79 substantial, and over 0.80 outstanding (Landis & Koch, 1977). In special education, the proposed minimal standards are IRA or interobserver agreement (IOA) = 80% and Kappa = 60% (Horner et al., 2005, p. 174). Thus, the agreement rate of 84% for study characteristics and 86% for quality indicators, and Kappa of 61% for quality indicators in the present study met the minimal standard.

Computing Effect Size and Design of Data Analysis

Effect size (ES) makes meta-analysis possible. In meta-analysis, ES is the dependent variable. ES allows for standardization of findings across studies so that they can be directly compared. Any standardized index such as standardized mean difference, correlation coefficient, or odds-ratio, can be considered as an ES as long as it is comparable across studies, represents the magnitude and direction of the relationship of interest (Lipsey & Wilson, 2001). There are many different types of ES indices; each is suited to different research situations. Each ES type may also have multiple methods of computation. To be comparable with each other and meaningful to analyze, the same ES statistics must be used for coding all the findings in a given meta-analysis (Lipsey & Wilson).

In this meta-analytic review, I adopted standardized mean differences ES to compute ES and multi-level hierarchical linear modeling (HLM) to analyze studies using each study's ESs. The rationale for choosing these statistics and analysis will

be discussed shortly. The following is the overview of meta-analysis steps (Beretvas & Paster, 2003; Lipsey & Wilson, 2001).

1. ES computation: transformation, adjustments, and outliers
2. Homogeneity analysis: HLM unconditional model
3. Moderator analysis: HLM conditional model
4. Aggregation between studies: Weighted ES, mean ES

ES Computation: Transformation, Adjustments, and Outliers

Transformation. For this review, the standardized mean difference is the ES statistic, which applies to research findings that compare the means of scores from outcome measures for experimental and control groups in studies with experimental design (e.g., ELL intervention vs. ELL control) or naturally occurring groups in studies employing quasi-experimental design (e.g., ELLs vs. L1 participants). This ES statistic is calculated from the statistical information reported in a study according to

$$\overline{ES} = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_{pooled}} \tag{1}$$

where X_{G1} is the mean for group 1, typically the experimental group, X_{G2} is the mean for group 2, typically the control group. S_{pooled} is the pooled standard deviation, defined as

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \tag{2}$$

where S^2_1 is the standard deviation for group 1, and S^2_2 is the standard deviation for group 2; n_1 is the number of subjects in group 1, and n_2 is the number of subjects in group 2.

ES statistics were not analyzed in their raw form, but adjusted to remove bias and outliers. Three procedures, including pretest effect size adjustment, correcting for small sample size bias, and removing outliers, were applied to the adjustments of the effect sizes. In the following sections, I describe each of the three procedures.

Effect size adjustment. The first procedure was adjustment of posttest ES for pretest ES. Whereas the focus of the present meta-analysis is posttest ES, some pretest ESs were obviously large (i.e., some studies did not achieve group comparability at pretest), particularly in studies employing quasi-experimental designs. A large pretest ES indicates serious pretreatment differences between treatment and control groups. Ignoring these pretests ESs would skew the distribution of posttest ESs; therefore, adjusting posttest effect sizes by subtracting pretest effect sizes from posttest effect sizes was applied to account for pretreatment differences (Hillocks, 1986; Wortman & Bryant, 1985; Yang, 2006). This adjustment was done for all studies that had a pretest ES.

Correction for small sample size bias. The standardized mean difference ES has been shown to have upward bias when sample sizes are small, especially for samples less than 20 (Hedges, 1981). To correct the small sample size bias, the following transformation procedure is applied.

$$ES'_{sm} = ES_{sm}\left[1 - \frac{3}{4N - 9}\right] \quad (3)$$

where ES_{sm} is the biased standardized mean difference ES statistics shown in Formula (1) above; and N is the total sample size, that is $n_1 + n_2$.

Removal of outliers. The last adjustment was to remove outliers, which means extreme values in the distributions of effect sizes. Outliers mean that there are studies that are notably different from others, which results in misrepresentation or distortion of the results (Hunter & Schmidt, 2004; Lipsey & Wilson, 2001). Outliers can be defined as ±2.00 or 3.00 *SD* from the mean ES (see Lipsey & Wilson, pp. 107-108). When outliers are found, a researcher first needs to make sure that the coding was accurate. When coding is determined to be correct, outliers can be removed (e.g., Graham & Perin, 2007; Swanson & Hoskyn, 1998; Swanson & Jerman, 2006; Swanson, Zheng, & Jerman, 2009) or adjusted to less extreme values (e.g., Shadish & Baldwin, 2005, Yang, 2006). Hunter & Schmidt stated that, in the case of meta-analysis methods in which sample sizes are often small to moderate, which is the case of this review, extreme values can occur simply because of large sampling errors and are not true outliers; therefore, the formula for sampling error variance used in meta-analysis, such as restricted maximum likelihood estimation, assumes and allows for such occasionally large sampling errors. Thus, they argued that elimination of such nonoutliers can result in overcorrection for sampling error and underestimation of *SD*, and most outliers should not be eliminated from the data. For this reason, they have generally not removed any but only the most extreme

outliers in conducting their meta-analyses.

Considering all these statistical concerns about outliers and the most common decisions made in previous meta-analyses in the area of intervention studies together, I decided to report all potential outliers in the Results section (Chapter IV), check each value for its possible reasons and remove only the most extreme as large as ± 3.00 SD. Then, I used restricted maximum likelihood (REML) as an estimation technique for HLM.

Homogeneity Analysis: HLM Unconditional Model

ESs can usually be assumed to be statistically independent. For this independence assumption, every study should have only one ES. In reality, studies often have multiple outcome measures. In fact, in terms of quality, a study is strongly recommended to have multiple outcome measures to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance (e.g., Gersten et al., 2005). However, in meta-analysis, multiple ESs due to multiple outcome measures violate the independence assumption. Multiple ESs could be reduced to a single ES by averaging all ESs or selecting only one ES (Lipsey & Wilson, 2001). HLM takes into account these dependences in the analysis, and the present meta-analysis used HLM.

Raudenbush and Bryk (2002) suggested that a meta-analysis could be viewed as a special case of the level-two multilevel analysis within HLM because meta-analytic data are hierarchically structured. HLM takes into account dependence of

multiple ESs within studies and heterogeneity between studies. When heterogeneity is assumed, a form of a random-effect model is recommended. If the model includes the addition of study-level characteristics or moderators to help explain some of the variability in ESs between studies, the random effect model is now called a mixed effect model, including both fixed and random effects. A mixed effect model is sought to model dependence of multiple ES estimates provided in a study while estimating how much ES estimates vary between studies (Beretvas & Pastor, 2003).

According to the basic concept of HLM, the two-level unconditional model can be expressed as follows. At level 1, the variability of ES parameter within a study is modeled so that

$$\zeta_j = \beta_0 + r_j \qquad u_j \sim N(0, \tau^2) \tag{4}$$

where β_0 represents the mean ES value for study j and r_j is the within-study error term assumed to be theoretically normally distributed with a mean of zero and a variance of τ.

Level 2 can be written as:

$$\beta_{oj} = \gamma_{00} + u_{oj}, \qquad u_{oj} \sim N(0, \omega^2) \tag{5}$$

where γ_{00} represents the overall mean ES for the population and u_{oj} represents the sampling variability between studies assumed to be normally distributed with a mean of zero and a variance of ω.

Thus, HLM can be used to test for homogeneity, traditionally conducted by Q statistics. Q statistics (Hedges & Olkin, 1985) are used to explore whether the observed variability in the distribution of effect size estimates is greater than would be expected from sampling error. If the Q statistic is significant, moderator analyses are warranted to identify the sources of this observed variability in the effect size distribution. The Q test of homogeneity has been found to have low power to detect heterogeneity of effects (Harwell, 1997), which means that although the test result may show ES estimates are not significantly different, this might be a result of poor power to detect true heterogeneity. In HLM, the unconditional model can be conducted to test for homogeneity. When the homogeneity assumption is rejected with the unconditional model, it means the distribution of ES is assumed to be heterogeneous. In other words, a single mean ES is not a good descriptor of the distribution; there are real within and/or between study differences. Therefore, a researcher needs to find moderators that influence ESs. HLM conditional models can achieve this goal.

Moderator Analysis: HLM Conditional Model

A more general mixed-effects model includes moderators or predictors in the level-two model, which is called a conditional model. In other words, the conditional multilevel meta-analysis is conducted to examine which study characteristics, or moderators, are affecting the strength of ESs. When a mixed-effects model has several moderators affecting ESs, the level-two model can be represented as:

$$\zeta_i = \beta_{0j} + \beta_{1j}X_{1ij} + r_{0ij}$$

$$\beta_{oj} = \gamma_{00} + \gamma_{1j}z_{1j} + u_{0j} \quad (6)$$

where X_{1ij} and Z_{1j} are moderators.

Besides dependency among multiple ESs due to multiple outcome constructs in a study, more subtle types of dependency may exist in meta-analysis. For example, some studies have two or more subsamples or group comparisons. In addition, more than two studies conducted by the same research team may be included in meta-analysis (Lipsey & Wilson, 2001). To handle these dependencies, one more level can be added as level 1 in HLM, resulting in a level-three model (e.g., Beretvas & Paster, 2003). However, Lipsey and Wilson stated that these dependencies are likely to be small in most applications and the standard practice in meta-analysis has been to define independence at the sample level (e.g., Kalaian & Raudenbush, 1996). In the present meta-analysis, besides considering the relatively minor concern about the independence among subsamples, some studies only have one comparison or subsample so that it does not have degrees of freedom in level-one of a possible level-three model; thus, I decided to adopt a level-two model.

To statistically model the dependencies in a level-two model, some statisticians prefer to employ "V-known" model or multivariate HLM (HMLM). To run the $\zeta =$ models, the intra-class correlation among ESs are estimated from the sample, deducted from published information, or imputed on the basis of past study (Kalaian

& Raudenbush, 1996). Unfortunately, few meta-analysis situations provide sufficient information that enables this method (Lipsey & Wilson, 2002). For this reason, the present review used a level-two model. As an estimation technique, restricted (or residual) maximum likelihood (REML) was used. The REML is a method for fitting linear mixed models. In contrast to conventional maximum likelihood estimation, REML can produce unbiased estimates of variance and covariance parameters, particularly when the number of level-two unit is not large enough.

Aggregation Between Studies: Weighted ES and Mean ES

The reason Gersten et al (2005) stressed the weighted ES in their criteria for evidence-based practice was that the weighted ES takes into account (a) the number of studies conducted on the specific intervention, (b) the number of participants in the studies, and (c) the magnitude and consistency of effects. Basically, "weighted" means weighting by the sample size of a study. Studies generally vary in size. A larger study is assumed to be a more "precise" estimate of true value. A simple approach is to weight each ES by its sample size; a better approach is to weight by the inverse variance because the standard error (SE) or variance is a direct index of ES precision. SE is used to create confidence intervals. The smaller the SE is, the more precise the ES is. The optimal weights for meta-analysis are:

$$w = \frac{1}{SE^2} \tag{7}$$

The SE in the standardized mean difference ES is

$$se = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{\overline{ES}_{sm}{}^2}{2(n_1 + n_2)}} \tag{8}$$

The following are the formulas for calculating mean ES, SE, Z-test, and 95% confidence interval with one example,

Mean ES:
$$\overline{ES} = \frac{\sum(w \times ES)}{\sum w} = \frac{41.82}{269.96} = 0.15 \tag{9}$$

SE of the mean ES:
$$se_{\overline{ES}} = \sqrt{\frac{1}{\sum w}} = \sqrt{\frac{1}{269.96}} = 0.061 \tag{10}$$

Z-test for the mean ES:
$$Z = \frac{\overline{ES}}{se_{\overline{ES}}} = \frac{0.15}{0.061} = 2.46 \tag{11}$$

95% confidence interval:
$$Lower = \overline{ES} - 1.96(se_{\overline{ES}}) = 0.15 - 1.96(.061) = 0.03$$
$$Upper = \overline{ES} + 1.96(se_{\overline{ES}}) = 0.15 + 1.96(.061) = 0.27 \tag{12}$$

CHAPTER IV

RESULTS

In this chapter, results are organized by the three research goals of this study, which were to determine (1) the availability of reading instruction for ELLs, (2) the effects of that instruction, and (3) which instructional programs can be considered evidence-based. First, with respect to availability, I present results of descriptive statistics categorized by study characteristics including intervention descriptors, ELL participant descriptors, outcome measure descriptors, and research design and publication year. Second, with respect to effects of instruction, I present results of the HLM unconditional and conditional models. The unconditional model produces overall effectiveness, and the conditional model produces the factors that affect the effectiveness. Last, to determine instruction that can be considered evidence-based, I present the weighted ESs from aggregated studies for each instructional program, and report the relation between study quality and ES.

Descriptive Analysis: Availability of Reading Instruction for ELLs

Studies Excluded and Included

As described in Chapter III, a search using electronic indexes, major journals, and ancestral tracking was conducted from January to May 2008. A total of 54 studies were identified. Among those, 25 studies were excluded with the following reasons (see Appendix D for the list of all excluded studies):

1. The study employed single-subject design (n = 16).

2. The study did not have a control group although it employed group experimental design (n = 4).

3. The study did not provide adequate data to compute the standard mean difference ES (n = 3).

4. The study did not include reading outcome measures of key components of reading (n = 2).

The remaining 29 studies were used for analyses. Because some studies included more than two comparisons, 44 independent samples or group comparisons from 29 studies were included. Furthermore, most studies included multiple outcome measures, which resulted in a total of 228 ESs. The number of samples and outcome measures are presented in Table 1. These 44 samples are the unit of analysis of level 2 in HLM; 35 samples for Data Set 1, and 9 samples for Data Set 2. The reason for dividing samples into two data sets will be described shortly.

Availability of Reading Instruction for ELLs

Table 1

Studies Included

Authors	No. of Comparisons	No. of ESs
Data Set 1: ELL vs. ELL		
Almaguer (2005)	1	3
Avila & Sadoski (1996)	1	2
Bernhard et al. (2006)	1	1
Calhoon, Al Otaiba, Greenberg et al. (2006)	1	3
Calhoon, Al Otaiba, Cihak et al. (2007)	2	8
Denton, Anthony, Parker, & Hasbrouck (2004)	2	6
Ehri, Dreyer, Flugman, & Gross (2007)	1	11
Giambo & KcKinney (2004)	1	2
Gunn, Biglan, Smolkowski & Ary (2000)	2	10
Kamps et al. (2007)	2	8
Leafstedt, Richards, & Gerber (2004)	3	12
McMaster, Kung, Han & Cao (2008)	1	8
Saenz, Fuchs, & Fuchs (2005)	4	12
Saunders & Goldenberg (1999)	3	6
Stuart (1999)	2	18
Stuart (2004)	1	6
Troia (2004)	1	6
Vaughn, Cirino et al. (2006)	2	28
Vaughn, Linan-Thompson et al. (2006)	1	12
Vaughn, Mathes et al. (2006)	1	12
Zhang & Schumm (2000)	2	4
Subtotal	35	178
Data Set 2: ELL vs. L1 or ELL At-Risk vs. ELL Not-at-Risk		
Chaippe, Siegel, & Wade-Woolley (2002)	2	16
Gerber et al. (2004)	2	10
Lesaux & Siegel (2003)	2	16
Proctor, Dalton & Grisham (2007)	1	2
Silverman (2007)	2	6
Subtotal	9	50
Total	44	228

Tables 2 to 5 present summaries of study characteristics categorized by intervention, participant, outcome measure, and design descriptors to show the availability of reading instruction for ELLs with respect to these variables. In addition, the availability of reading instruction by combined variables of instructional tier, grade, and instructional components are presented in Table 6.

Table 2
Characteristics of Intervention (N=44)

Descriptors		*n*	*(%)*
Language			
	English	38	(86.4)
	Spanish	5	(11.4)
	English+Spanish	1	(2.3)
Instructional Components Included			
	Phonemic awareness	27	(61.4)
	Phonics	24	(54.5)
	Fluency	20	(45.5)
	Vocabulary	12	(27.3)
	Comprehension	24	(54.5)
Multiple Instructional Components			
	1 component	14	(31.8)
	2 components	14	(31.8)
	3 components	4	(9.1)
	4 components	7	(15.9)
	5 components	5	(11.4)
Tiers			
	Tier 1	27	(61.4)
	Tier 2	17	(38.6)
Fidelity			
	Evaluated	24	(54.5)
	Not evaluated	20	(45.5)

Table 2 (Continued)

Descriptors		*n*	(%)
Session Length			
	45 min or less	34	(77.3)
	Over 45 min	7	(15.9)
	Not specified	3	(6.8)
Session Per Week			
	3	13	(29.5)
	4	5	(11.4)
	5	12	(27.3)
	Not specified	14	(31.8)
Total Duration			
	12 wks or less	8	(18.2)
	13-24 wks	11	(25.0)
	25-72 wks	14	(31.8)
	Not specified	11	(25.0)
Grouping			
	One to one tutor	3	(6.8)
	Pair	9	(20.5)
	Small group	16	(36.4)
	Whole class	16	(36.4)
Interventionist			
	Classroom teacher	34	(77.3)
	Researcher	3	(6.8)
	Under-/graduate student	4	(9.1)
	Hired teacher	3	(6.8)

Table 3

Characteristics of Participants (N=44)

Descriptors	n	*(%)*
Grade Included		
Preschool	4	(9.1)
Kindergarten	14	(31.8)
Grade 1	20	(45.5)
Grade 2	6	(13.6)
Grade 3	10	(22.7)
Grade 4	11	(25.0)
Grade 5	13	(29.5)
Grade 6	5	(11.4)
Grade: Single vs. Multiple		
Single grade	28	(63.6)
Multiple grades	16	(36.4)
Grade: Beginner vs. Upper		
Beginner (pre-grade 2)	32	(72.7)
Upper (grade 3-6)	12	(27.3)
ELL Language		
Spanish	32	(72.7)
Mixed language	9	(20.5)
Not specified	3	(6.8)
ELL Percentage in the Classroom		
100% ELL	30	(68.2)
Less than 100% ELL	14	(31.8)
LEP by Definition		
100% LEP	29	(65.9)
Less than 100% LEP	6	(13.6)
Not specified	9	(20.5)
ELL Reading Performance		
Low or at-risk	18	(40.9)
Average or higher	5	(11.4)
Full range	17	(38.6)
Special Education	2	(4.5)
Not specified	2	(4.5)
ELL English Language Proficiency		
Low	30	(68.2)
Full range	1	(2.3)
Not specified	13	(29.5)

Table 4

Characteristics of Outcome Measures (N=44)

Descriptors		n	*(%)*
Outcome Construct			
	Phonemic Awareness	24	(54.5)
	Phonics	26	(59.1)
	Fluency	21	(47.7)
	Vocaburary	13	(29.5)
	Comprehension	28	(63.6)
Number of Outcome Construct Included			
	1	8	(18.2)
	2	18	(40.9)
	3	8	(18.2)
	4	6	(13.6)
	5	4	(9.1)
Score: Standardized vs. Raw			
	Stadardized score only	9	(20.5)
	Raw score only	19	(43.2)
	Both	16	(36.4)
*Measures Aligned with Instructional Components**			
	Aligned	29	(65.9)
	Not alinged	1	(2.3)
	Both	14	(31.8)
Timing of Measurement			
	Immediate	40	(90.9)
	Follow-up	4	(9.1)

*Note: Aligned means that the constructs of the outcome measures match the instructional components of the intervention in the study (e.g., a study including a phonics instructional component and phonics outcome measure).

Table 5

Characteristics of Research Design and Publication Year(N=44)

Descriptors	n	*(%)*
Design		
Experimental	25	(56.8)
Quasi-experimental	19	(43.2)
Type of Comparision		
ELL vs. ELL	35	(79.5)
ELL vs. L1 students	7	(15.9)
ELL at-Risk vs. ELL not-at-risk	2	(4.5)
Publication Year		
1996	1	(2.3)
1999	5	(11.4)
2000	3	(6.8)
2002	3	(6.8)
2003	2	(4.5)
2004	10	(22.7)
2005	5	(11.4)
2006	6	(13.6)
2007	8	(18.2)
2008	1	(2.3)

Table 6
Availability of Instruction by Tier × Grade

Tier	Grade	Instructional Components included	*n*
Tier 1 (*n=27*)	Beginner: PreK-2 (*n=16*)	Phonemic Awareness	14/16
		Phonics	13/16
		Fluency	5/16
		Vocabulary	2/16
		Comprehension	5/16
	Upper: 3-6 (*n=11*)	Phonemic Awareness	0/11
		Phonics	0/11
		Fluency	5/11
		Vocabulary	3/11
		Comprehension	7/11
Tier 2 (*n=17*)	Beginner: PreK-2 (*n=16*)	Phonemic Awareness	13/16
		Phonics	11/16
		Fluency	10/16
		Vocabulary	7/16
		Comprehension	11/16
	Upper: 3-6 (*n=1*)	Phonemic Awareness	0/1
		Phonics	0/1
		Fluency	0/1
		Vocabulary	0/1
		Comprehension	1/1

HLM Analysis: Effects of Reading Instruction for ELLs

Adjustment of Effect Size

The standard mean difference ES statistics were adjusted to remove bias and outliers using three procedures: pretest effect size adjustment for studies that had pretest scores, correction for small sample size bias, and removal of outliers. Among 228 ESs, 156 ESs (or 68.4%) were adjusted based on pretest ES. Specifically, the percentages of ESs adjusted for pretest ES were 88.2% (30/34), 70.6% (72/102), 50.0% (15/30), 66.7% (12/18), and 61.4% (27/44) for phonemic awareness, phonics, fluency, vocabulary, and comprehension outcome measures, respectively. Then, all ESs were corrected for small sample size bias. Removal of outliers is described in more detail as part of the HLM analysis results.

Two Different Kinds of Data Sets

Before describing the HLM analysis results, it is important to distinguish instances in which researchers made two different kinds of comparisons to examine the effect of instructional programs: (1) a comparison between ELL intervention group and ELL control group, and (2) a comparison between ELLs and L1 students or between at-risk ELLs and not-at-risk ELLs who received the same instruction (Rolstad et al, 2006). In the first kind of comparison, a researcher examines if the instruction is effective, and seeks to reject the null hypothesis; in contrast, in the

second comparison, a researcher investigates whether the instruction designed for L1 has no differential effect for ELL groups, and seeks to fail to reject the null hypothesis. In other words, the researcher expects relatively large ESs in favor of intervention group for the first kind of comparison, and near zero ESs for the second kind. In addition, whereas most studies of the first kind may employ experimental designs, the second kind of comparison is typically quasi-experimental.

In this study, Data Set 1 consists of studies comparing ELL treatment groups to ELL control groups, and includes 35 samples and 178 ESs; and Data Set 2 consists of studies comparing ELLs to L1 students or at-risk ELLs to not-at-risk ELLs, and includes 9 samples and 50 ESs. Due to different expectations about the magnitude of ESs and, subsequently, the different goals in the hypothesis testing, the following analyses are conducted and reported separately by these two data sets.

Data Set 1: Comparison between ELL Treatment and ELL Control Groups

To remove outliers, extreme ESs were checked in the distribution of ESs by using SPSS stem-and-leaf plots, normal Q-Q plots and boxplots. Among 178 ESs, 8 potential outliers were found: 7 ESs larger than +1.7 *SD* and 1 ES less than -1.6 *SD*. Table 7 shows the potential outliers and the decision rule for removal. Close examination of these outliers revealed that two extreme ESs were caused by severe nonequivalence at pretest and one ES showed a significant discrepancy between its value calculated by the ES formula and the one reported in the original article. The *Z*

Table 7
Potential Outliers and Decision of Removal

author	sample	outcome measure	pre ES	post ES	Es adj. by preES	N	ES adj. by smallN	Z score	Decision
Kamps et al. (2007)	grade 1	phonics (WAT[1])		2.84	2.84	88	2.82	3.79	remove
Kamps et al. (2007)	grade 1	phonics(WID[2])		2.08	2.08	88	2.07	2.59	keep
Kamps et al. (2007)	grade 2	phonics (WID)		1.91	1.91	88	1.88	2.30	keep
Kamps et al. (2007)	grade 2	comprehension		1.78	1.78	55	1.76	2.10	keep
Leafstedt et al (2004)	low ability	phonics(WID)	-3.37	1.40	4.76	18	4.54	6.52	remove
Leafstedt et al (2004)	high ability	phonics (WAT)	3.88	2.17	-1.72	12	-1.59	3.22	remove
Saenz et al (2005)	high ability	fluency(WC[3])	-0.22	1.65	1.88	12	1.73	2.06	keep
Zhang et al (2000)	n/a	vocabulary		1.75	1.75	40	1.72	1.72	keep

Note. WAT[1]=word attack;WID[2]=word identification;WC3=The Comprehensive Reading Assessment Battery Word Correct Score

scores of these extreme ES were larger than +3.00 or smaller than -3.00. As a result of removing three outliers, 175 ESs from 35 comparisons were used for HLM.

Unconditional meta-analysis: Homogeneity analysis. An unconditional meta-analysis with the method of restricted maximum likelihood estimation was used to estimate the overall mean ESs and to examine the variability among samples. The results are shown in Table 8.

Table 8

Results for the Unconditional Model Analysis of 35 Samples

	Fixed Effect				
	Coefficient	*SE*	*t* Ratio(*df*)	95% CI	
				Lower	Upper
Intercept	.50	.07	7.15**(34)	.37	.63
	Random Effect				
	Variance Component	*SD*	*Chi-square*		
Intercept	.136	.368	179.17**		
Level-1	.140	.374			

Note. **$p < .01$, *SE*=Standard Error, CI=Confidence Interval, *SD*=Standard Deviation

The intercept in the fixed effect is the overall mean ES from 35 samples including 175 ESs. The intercept was 0.50 (t =7.15, p<.01), meaning that on average, the effect of reading instruction for ELLs is moderate with a mean ES of 0.50. The effect size, Cohen's *d*, is defined as "small, *d* = .2," "medium, *d* = .5," and "large, *d* = .8" (Cohen, 1988). For interpretation, an ES of 0.0 indicates that the mean of the treated group is at the 50th percentile of the untreated group. An ES of 0.2 (small) .08" (Cohen, 1988). For interpretation, an ES of 0.0 indicates that the mean of the treated group is at the 50th percentile of the untreated group. AN ES of 0.2 (small) indicates the mean of the treated group is at the 58th percentile of the untreated group, 0.5 (medium) indicates at the 67th percentile, 0.8 (large) indicates at the 79th percentile. An ES of 1.0 indicates at the 84 percentile, and 2.0 indicates at the 97.7 percentile.

The variance component that provides information about variability among samples was 0.136. Thus, 49% of the total variance in ESs (0.136 / (0.136+0.140)) was attributable to the between-sample level. The magnitude of the variance component, 0.136, was significant (chi-square=179.17, $p<.01$), which indicates the variability was not attributable to sampling error alone. Thus, an additional model with predictors that account for the variability is needed (Rosenthal, Hoyt, Ferrin, Miller, & Cohen, 2006).

Conditional meta-analysis: Moderator analysis. A conditional meta-analysis was conducted to examine variability among samples and identify factors affecting the strength of the effect. The within-sample and between-sample level predictive models were fitted for selected predictors, and are shown in Table 9. These predictors were selected because of their potential significance and because there were no missing data for these particular variables. For ease of interpretation, all but three predictors were dummy-coded with 0 for the reference condition and 1 for the comparison and no-centering was the choice for the model. Exceptions were the outcome constructs, which were coded as categorical data (1 for phonemic awareness, 2 for phonics, 3 for fluency, 4 for vocabulary, 5 for comprehension); quality indicator scores; and the number of instructional components (0 to 4 in the form of {the number of instructional components -1} because this is ordinal data).

Table 9

Predictors Selected for Moderator Analysis

Predictors		Recoding As a Dummy Variable	
		0	1
Within-Sample Level			
Outcome Measures			
	Standard score	standard	raw
	Align with instruction	aligned	not aligned
	Type of construct*	1 to 5	
Between-Sample Level			
Interevention			
	Language	English	Spanish
	Tier	1	2
	Session length	45 min or less	more than 45 min
	No. of components*	0 to 4	
ELL Participants			
	Grade level	beginner (pre-2nd)	upper (3-6th)
	Grade single	single	multiple
	Home language	Spanish	not Spanish
	SES	low	average or higher
	English proficiency	low or at-risk	average or higher
Outcome Measures			
	Timing of measurement	immediate	follow-up
Design and Quality			
	Design	experimental	quasi-experimental
	Quality score*	4 to 10	

* predictors that were not coded with 0 and 1

The results of the conditional meta-analysis are presented in Table 10. For outcome measures at the within-sample level, whether the outcome was a standard or raw score and whether the construct was aligned with the instructional component did not lead to significant differences in mean ESs. The coefficient for the type of outcome construct was significant (-0.07, p= 0.021), meaning that ESs decreased with the increase of codes. The code started with 1 for phonemic awareness and ended with 5 for comprehension. The weighted ESs were 0.41 for phonemic awareness (n=26), 0.33 for phonics (n=72), 0.38 for fluency (n=27), 0.34 for vocabulary (n=11), and 0.32 for comprehension measures (n=39). Thus, among 5 outcome measures, phonemic awareness instruction appears to have the highest ES.

At the between-sample level, with the other predictors held constant, the significant coefficients means that the mean ESs are significantly higher for studies in the former, reference condition than those in latter, comparison condition: specifically, for language of intervention (-0.29, p=.016), when the intervention language was English, ES=.46, and when the intervention language is Spanish, ES=0.18; for session length (-0.22, p=0.003), when a session lasted 45 min or less, ES=0.52, and when a session lasted more than 45 min, ES=0.21; for beginner and upper grade (0.71, p=0.010), when ELL participants in 3rd to 6th grade were included, ES=0.58 and when the participants were in preschool to 2nd grades, ES=0.33; for single and multiple grades included (-1.23, p=0.000), when ELL

participants included were in a single grade, ES=0.36, and when multiple grades were included, ES=0.29; and for SES (0.64, p=0.001), when SES of ELL participants was average or higher, ES=0.46, and when SES was low, ES = 0.31 In addition, with the other predictors held constant, the mean ESs decreased significantly with the increase of the quality of studies (-0.18, *p*=.018). The other variables, such as research design, tier of instruction, the number of instructional components included, ELL students' home language and their English language proficiency were not related to ES.

Additionally, the relation between each instructional component and ES was analyzed separately because of its practical implication. As shown in Table 11, none of the instructional components were significant predictors of ES. In these HLM analyses, the instructional component was coded as 1 for included and 0 for not included, so negative coefficient values mean the ESs decreased when the component was included.

Table 10
Results for the Conditional Models Analysis of 35 samples

Fixed Effect	coefficient	SE	T-ratio	Approx. d.f.	*p*-value
Within-Sample Level					
Outcome Measures					
Standard score	-0.14	0.10	-1.41	34	0.168
Align with instruction	-0.07	0.11	-0.54	34	0.591
Type of construct*	-0.07	0.03	-2.43	34	0.021
Between-Sample Level					
Interevention					
Language	-0.29	0.11	-2.62	22	0.016
Tier	0.14	0.18	0.77	22	0.452
Session length	-0.22	0.07	-3.39	22	0.003
No. of components*	0.09	0.05	1.89	22	0.071
ELL Participants					
Grade level	0.71	0.25	2.85	22	0.010
Grade single	-1.23	0.14	-9.08	22	0.000
Home language	-0.11	0.10	-1.07	22	0.295
SES	0.64	0.15	4.12	22	0.001
English proficiency	0.08	0.17	0.49	22	0.625
Outcome Measures					
Timing of measurement	0.20	0.11	1.71	22	0.101
Design and Quality					
Design	-0.29	0.25	-1.15	22	0.262
Quality score*	-0.18	0.07	-2.57	22	0.018

* predictors that were not coded with 0 and 1

Table 11
Results for the Conditional Models Analysis for Instructional Components

Fixed Effect	coefficient	SE	T-ratio	Approx. d.f.	*p*-value
Instructional Component Included					
Phonemic Awareness	-0.08	0.18	-0.34	29	0.739
Phonics	-0.17	0.25	-0.69	29	0.496
Fluency	0.14	0.19	0.75	29	0.457
Vocabulary	0.00	0.17	0.-0.00	29	0.998
Comprehension	-0.19	0.21	-0.92	29	0.363

Data Set 2: Comparison between ELL and L1 Receiving the Same Instruction

Data Set 2 consists of studies comparing ELLs and L1 students or at-risk ELLs and not-at-risk ELLs receiving the same instruction. In these studies, a primary research question is whether the instruction works as well for ELLs or at-risk students as for L1 or not-at-risk students. In other words, the researchers do not expect ES differences between groups; thus, they seek to fail to reject the null hypothesis. This is radically different from the conventional approach in which researchers seek to reject the null hypothesis, as seen in studies comparing ELLs in intervention and ELLs in control groups (Rolstad, Mahoney, & Glass, 2005).

Because of this unusual approach in which the researchers seek to fail to reject the null hypothesis, not many studies with this comparison have been

attempted. Indeed, Data Set 2 was much smaller, with 9 samples and 50 ESs, compared to Data Set 1, with 35 samples and 175 ESs. However, the practical implication of this comparison is not negligible, in particular in the context of RTI, in which tier 1 instruction targets all students in a classroom, including ELLs and at-risk students as well as L1 and not-at-risk students. Thus, it is useful to know whether instruction has different effects for different groups of students. However, it should be noted that, just as conventional hypothesis testing has a possibility of rejecting the null hypothesis with 5% error rate when the hypothesis is true, which is called as Type I error, α, or false positive, this unconventional hypothesis testing also has a possibility of failing to reject null hypothesis with 5% error rate when the hypothesis is false, which is Type II error, β, or false negative.

The overall mean effect of instructional programs in Data Set 2 was 0.07 (*SD*=0.54), and the 95% confidence interval with the Standard Error (SE) of 0.076 was -0.08 to 0.22. Thus, this ES is not significantly different from zero, which means that, on average, the instructional programs in Data Set 2 do not produce differential effects for ELL or at-risk students in studies as they do for L1 and not-at-risk students. In this data set, when the average ES is positive it means that the reference group (i.e., ELL or at-risk students) outperform the comparison group (i.e, L1 or not-at-risk students). Unfortunately, the number of level-two units (*n*=9) was not large enough to conduct level-two HLM to identify factors moderating these effects.

Identification of Evidence-Based Instructional Programs

The third primary goal of this study was to determine which instructional programs can be considered evidence based using Gersten et al.'s (2005) criteria of at least two high quality studies and a weighted ES significantly greater than zero. The criteria for considering as *promising* are at least two high quality studies and a 20% confidence interval for the weighted effect size that is greater than zero. For this meta-analysis, all studies were evaluated in terms of the quality indicators proposed by Gersten et al. . Again, due to different expectations for the magnitude of ESs, the following analyses were conducted and reported separately by Data Sets 1 and 2.

Data Set 1: Comparison between ELL Treatment and ELL Control Groups

Table 12 presents the frequency of quality indicator scores for the 35 samples in Data Set 1. The score was calculated only for essential quality indicators, although the evidence-based criteria proposed by Gersten et al. (2005) include 10 essential and 7 desirable qualities. Because of the differential weight issue for essential and desirable quality indicators (Lane, Kalberg, & Shepcaro, 2009; Cook et al, 2009), it might be problematic to give one credit for each component in desirable quality indicators. In specific, in Gersten et al's (2005) criteria, be considered acceptable,

Table 12
Frequency of Quality Indicator Scores (N=35)

Quality Scores (out of 10)	Number of Studies	(%)
4	1	(2.9)
6	2	(5.7)
7	7	(20.0)
8	8	(22.9)
9	8	(22.9)
10	9	(25.7)
Mean	*SD*	
8.31	1.45	

a study should need to meet 9 out of 10 essential quality indicators but need only 1 out of 8 desirable quality indicators. To be considered high quality, a study would need to meet four desirable quality indicators as well as 9 essential quality indicators. Thus, it is reasonable to make a distinction between two indicators, with the possibility of the higher weight for the essential quality indicators, and lower weight for the desirable quality indicators. The relative weights can be determined by conducting empirical studies by using both indicators. None of the reviews in the recent special issue of *Exceptional Children* devoted to use evidence-based practice guideline including quality indicators used both indicators.

Regarding the use of both indicators, there is one exception in previous

research. Test, Fowler, Brewer, & Wood (2005) assessed the presence of the QIs that Gersten et al. (2005) proposed. They assessed 23 criteria, examining essential and desirable QIs together and including criteria regarding the conceptualization of a study that Gersten et al. included in their QIs for research proposals. However, no distinction between essential and desirable quality indicators in Test et al.'s could be problematic (Cook et al, 2009). Cook et al. suggested additional field tests are needed to make a logical distinction between these two kinds of indicators.

The Pearson correlation between quality indicator scores and ESs for 35 studies was marginally significant, r =-0.331 (p=0.052). It means that studies with higher quality tended to have lower ESs. This association between lower quality and higher effect size is supported by the decreased correlation after removing 3 extreme ESs from studies that seem to have lower quality, such as non-equivalence at pretests; the correlation become less strong, r=-0.264 (p=0.126).

To calculate the weighted ES for a specific instructional program, it is necessary to aggregate the studies conducted on the specific program (Gersten et al., 2005). Table 13 presents summaries of 35 studies conducted on 13 different instructional programs along with their ESs and quality scores. Based on the criteria for evidence-based practice proposed by Gersten et al., three instructional programs were identified as evidence-based or promising; these are presented in Table 14. The Keyword method can be considered an evidence-based practice, and Peer-Assisted

Table 13

Summaries of Studies Comparing ELLs intervention and ELLs Control

Study	Intervention	language of INT	components[7] (PA[1],p,f,v,c)	Intervention: length frequency,duration,grouping,interventionist	Tier	Grade	Participants: N[10](Int[11]/Con[12]), Language,Ability,SES[13]	ES by sample (if not specified, ELvsEL)		ES	QI[15] Score
Almaguer(2005)	DyadReading	E[5]	f	30min,5/wk,9wks,pair,teacher	1	3	80(40/40),Spanish,GE[14],lowSES			0.39	9
Bernhard et al.(2006)	EarlyAuthorsProgram	E	PA,p	unk,1yr,whole,specialists+teachers	1	p[8]	68(51/17),Mixed,GE,lowSES			0.80	4
Leafstedt et al. (2004)	EarlyReadingProject(ERP)	E	PA	30min,10session,grof3-5,researcher	2	k[9]	62(16/46), Spanish,full,lowSES	High Middle Low	0.43 1.19 0.47	0.70	7
Troia(2004)	FastForWord	E	PA,p,c	100min,5/wk,4wks,whole,Coach	1	1-6	191(99/92),Spanish,full+SE,NS			0.06	8
Stuart(1999)	JollyPhonics	E	PA,p	60min,5/wk,12wks,whole,teacher	1	P	112(55/57),Sylheti,GE,lowSES	immedeate 1yr f-up	0.12 0.25	0.29	7
Stuart(2004)							101,Sylheti,GE,lowSES	2yr f-up	0.51		7
Avila(1996)	KeyWord	S[6]	v	30min,3sessions,whole,teacher	1	5	63(30/33),Spanishc,low,lowSES			0.88	9
Zhang(2000)		E/S	v	30min,3sessions,grof10,researcher	1	5	60(20/20/20),Spanish,GE,NS	English Int Spanish Int	1.44 1.27	1.35	9
Saunders(1999)	Log&Coversation	E	c	45min,2lessons/day,2days(comGr=4days),teacher	1	4-5	116,Spanish,full,NS	Log Conversatio Combined	-0.25 0.41 0.81	0.32	8

Table 13 (Continued)

Study	Intervention	language of INT	components (PA,p,f,v,c)	Intervention: length frequency,duration,grouping,interventionist	Tier	Grade	Participants: N(Int/Con), Language,Ability,SES	ES by sample (if not specified, ELvsEL)		ES	QI Score
Giambo(2004)	PA program[1]	E	PA	20min,3/wk,19wks,grof 5,teacher	1	k	80(40/40),Spanish,full,lowSES			0.18	8
Saenz et al (2005)		E	f,c	30min,3/wk,15 wks, teacher	1	3-6	119(59/60), Spanish,,full+LD,NS	LD Low Ave High	0.61 0.23 0.27 0.86	0.49	9
Calhoon et al (2006)	PALS	E	PA,p,f,c	30-35min,3/wk,20 wks,pair,teacher	1	1	78(41/37), 49Spanish(25/24); 29 L1(16/13), full+6SE,lowSES			0.61	10
Calhoon et al (2007)		E	PA,p,f,c	30-35min,3/wk,20 wks,pair,teacher	1	1	76(43/33), 60Spanish(33/27) 16L1(10/6),full+11SE,lowSES	EL+L1vs.EL+ L1	0.27 0.63	0.45	10
McMaster et al.(2008)		E	PA, p,f	30min,4/wk,18wks,pair, teacher	1	K	40(20/20),Mixed,full,lowSES			0.32	10
Vaughn&Linan-Thomson et al.(2006)		S			2	1	64(31/33),Spanish,at-risk,lowSES			-0.08	10
Vaughn&Mathes et	Proative Reading	E	PA,p,f,c	50min,5/wk,7months,group of3-5,teacher	2	1	41(22/19),Spanish,at-risk,lowSES			0.56	10
Vaughn&Cirino et al.(2006)		E			2	1	91(43/48),Spanish,at- risk,lowSES			0.11	10
		S			2	1	80 (35/45),Spanish,at-risk,lowSES			0.30	10

Table 13 (Continued)

Study	Intervention	language of INT	components (PA,p,f,v,c)	Intervention: length frequency,duration,grouping,interventionist	Tier	Grade	Participants: N(Int/Con), Language,Ability,SES	ES by sample (if not specified, ELvsEL)		ES	QI Score
Gunn et al. (2000)	ReadingMastery+Corrective Reading	E	PA, p,f,c	25-30min,5/wk, 15mo(2yr,Time3), grof 2-3, Assistant	2	k-3	256, Spanish+non-Spanish,low,NS	immedeate f-up	0.45 0.37	0.41	7 8
Ehrietal(2007)	ReadingRescue(RES)	E	PA,p,f,v,c	30-40min,5/wk,26wks,tutor (pro+para)	2	1	126(64/62),Spanish,at-risk,lowSES			0.52	10
Denton (2004)	Read Well	E	p,v,c	40min,3/wk,10wks,tutoring,undergraduate	2	2-5	33(19/14), Spanish,AtRisk,NS			0.26	8
	Read Naturally	E	f,v,c	40min,3/wk,10wks, tutoring,undergraduate	2	2-5	60(32/28), Spanish,AtRisk,NS			0.00	8
Kamps et al.(2007)	RW[2]+DI[3] RN[4]+DI	E	PA,p,f,c	NS,1yrs,grof3-6,teacher	2 2	1 2	144(84/60),Spanish,non-responder,fullSES	Grade1 Grade2		1.73 1.41	6

Note : PA progrma[1]=phonological awareness intervention program; RW[2]=Read Well; DI[3]=Direct Instruction; RN[4]=Read Naturally; E[5]=English, S[6]=Spanish; Components[7]=(p=phonics, f=fluncy, v=vocabulary, c=comprehension); p[8]=preschool; k[9]=kindergarten; N[10]=total sa

Table 14

Evidence-Based or Promising Reading Instruction for ELLs in Elementary Grades

Intervention	No. of Study	language of intervention	instructional components (PA,p,f,v,c)	Grade	Quality Score	Weighted ES	SE[1]	Z	CI[2]	Note
KeyWord	2	E,S	v	5	9,9	1.07	0.20	5.29*	.67~1.47	Evidence-Based Practice
PeerAssisted LearningStrategies (PALS)	4	E	f,c PA,p,f,c PA,p,f	3-6 1 K	9,10,10,10	0.57	0.33	1.73,n.s.	-.08 ~1.12	promising
Proactive Reading	2	E	PA,p,f,c	1	10,10	0.35	0.05	7.17*	.25~.45	Evidence-Based Practice
	2	S	PA,p,f,c	2	10,10	0.04	0.05	0.76,n.s.	-.06~.14	not Evidence-Based Practice

Note: SE[1]=Standard error; CI[2]=95% of Confidence Interval

Table 15

Summaries of Studies Comparing ELLs and L1 Students in the Same Intervention

Study	Intervention	language of INT	components (PA,p,f,v,c)	Intervention:length frequency,duration,grouping,Interventionist	Tier	Grade	Participants N(EL/L1), language,ability,SES	ES by sample (if not specified, ELvsL1)		WES	Quality Score
Gerber et al. (2004)	Core Intervention Model	S	PA	30min,10session, grof3-5, bilingual undergrats	2	k	82kEL(T37at-risk/C45 not-at-risk)-> 43Gr1(T28/C15) ,Spanish,At-RiskReaders, NoSESinfo	kindergarten	0.45	0.18	6
		S/E	p		2	1		Grade 1	-0.09		
Chiappe et al. (2002)	Lauch into Reading Success	E	PA,p	20min,3-4/wk, 2yrs,classroom& resource teachers	1	k,1	857 (118/739) ; 140AtRisk(32/108)+717NotAtRisk (86/631),GE,middleSES	at-risk	0.23	0.14	6
								Not-at-risk	0.04		
Lesaux & Siegel (2003)		E	PA for K, pforGr1	20min,3-4/wk, 2yrs,classroom& resource teachers	1	k,1	1162K(866NotAtRisk(100/766)+ 296AtRisk(236/60),Mixed,GE, FullSES	RD	-0.17	-0.01	6
								Ave	0.15		
Silverman (2007)	Multidimensional Vocabulary Program	E	voc	30-45min, 3/wk,14wks, whole,teacher	1	k	72(44/28), Mixed,Half, lowSES	immediate	-0.10	-0.10	9
								follow-up	-0.09		
Proctor et al. (2007)	Universal Literacy Environment	E	voc,comp	45min,3/wk,4wks,Computer,Tec& Co-teachers	2	4	30(16/14),Spanish,atRisk, lowSES			0.16	5

Learning Strategies (PALS) and Proactive Reading can be considered promising practices.

Data Set 2: Comparison between ELL and L1 Intervention Groups

Table 15 presents summaries of 9 studies conducted on 5 different instructional programs along with their weighted ESs and quality scores. None of the programs was supported by the first criterion of evidence-based practice, which is at least two high quality studies.

CHAPTER FIVE

DISCUSSION

ELLs have as high a potential to learn as their native English-speaking peers. Congruent with this belief, the NCLB sets the same high academic expectations for ELLs as for native English speaking students. However, the achievement gaps between ELLs and L1 students, including gaps in reading test scores, have been significant and persistent. To reach their potential and close these gaps, ELLs need quality, effective instruction based on empirical evidence.

The primary purpose of this review was to identify the available, effective practices in reading instruction for ELLs, and to determine which of those practices can be considered evidence based. The findings from this meta-analysis demonstrate that there are about 13 instructional programs targeting ELL students that have been empirically examined using experimental or quasi-experimental designs, and the overall effect is moderate with a mean ES of 0.50. However, among those instructional practices, only a few were identified as evidence-based or promising based on the criteria proposed by Gersten et al., (2005).

In this chapter, major findings are summarized and discussed with their implications for research, practice, and policy. Before summarizing and discussing the findings, it is important to (a) identify study limitations and (b) determine how this review differs from previous reviews. Specifically, studies included in this review are

compared to two previous reviews conducted by Cheung and Slavin (2005) and Klingner et al. (2006), and also to a recent review released by What Work Clearinghouse (WWC, 2009).

Limitations

Before presenting major findings of this review, there are several limitations that must be considered. First, the review was limited to experimental and quasi-experimental studies. Although studies with single-subject designs might contribute significantly to our understanding of how to teach ELLs in reading, ES statistics for studies using this design are different from those of studies with group designs. Moreover, qualitative studies were not included in the present review. Although multiple meta-analyses combining studies using different designs can be performed on the same body of literature (Lipsey & Wilson, 2001), this approach has rarely been used in the literature, mainly because very few studies provide sufficient information to convert and combine effect sizes (DeCoster, 2004).

Second, and related to first limitation, this review was limited to studies published in peer-reviewed journals. Unpublished research, such as dissertation, reports and conference presentations, was not included. Thus, potential publication bias could result in upwardly biased effect sizes. The nature and impact of publication bias will be discussed in greater detail shortly.

Third, because this review included studies focused only on the five key instructional components identified by The National Reading Panel (NICHD, 2000; phonemic awareness, phonics, fluency, vocabulary, and text comprehension), this review cannot draw conclusions about other instructional components, such as oral language proficiency in English. In fact, the National Literacy Panel on Language-Minority Children and Youth (2006) and Helman and Burns (2008) recommended that reading programs for ELLs should include intensive English language development as well as instruction in five key components. However, there were only a few studies that integrated oral language development activities into literacy skills instruction for ELLs. Instead, n the case of the panel, the recommendation was a hypothesis drawn from several other findings in the report. For this present review, oral language development activities was not included as the instructional components, and two studies that used only oral language development measures without outcomes on key components were excluded. .

Fourth, related to the limitation above, vocabulary instruction has been the focus of less research than the other components. Whereas more than 20 studies focused on phonemic awareness, phonics, fluency, and comprehension, only 13 focused on vocabulary instruction. Thus, the results of vocabulary instruction are tentative and less conclusive, even though meta-analysis accounts for the importance of sample size by adjusting for sample size at various points in the analysis, such as pretest ES adjustment, small sample size bias correction, and inverse variance

weight. This tentativeness should also be applied to the results of other variables that were represented by a relatively small number of effect sizes, such as Spanish as the language of instruction, which had only four studies.

Fifth, as in most meta-analytic reviews, I had to make many decisions in terms of inclusion criteria, the type of ES adjustment, removal of outliers, levels of and predictors in the HLM, all of which may be criticized for subjectivity. Although this kind of decision is necessary for conducting any meta-analysis, other researchers might question one or more of the decisions I made. For this reason, I tried to make my decision-making as reasonable and transparent as possible.

How this Review Differs from Previous Reviews

A total of 44 studies that focused on elementary reading for ELLs were included in the present meta-analytic review. This number is certainly quite encouraging when considering the relatively small numbers of studies identified in previous reviews: 13 studies in Cheung and Slavin (2005) and 7 studies in Klingner et al. (2006). Aside from the fact that the current study included studies published since the previous reviews, the most plausible reason for the differences in the number of studies included is the different inclusion criteria adopted.

Whereas Cheung and Slavin (2005) included only studies employing random assignment or matching and excluded studies with pretest differences of

more than one standard deviation, the present study included both experimental and quasi-experimental studies, and did not exclude studies with pretest differences of more than one standard deviation; instead, in this review, the pretest differences were subtracted from the posttest differences, and only three outcome measures having more than three standard deviations were excluded. In addition, while Cheung and Slavin included studies involving elementary (K–6) children, the present review included four studies conducted for preschool children. The reason for including studies on preschool children was that theories on second language acquisition address the need of instructional attention to the emergent stage, which corresponds to preschool-kindergarten level, with its emphasis on phonemic awareness and phonics instruction. Also, while Cheung and Slavin included studies with minimum treatment duration of 12 weeks, the present review included eight studies with treatment duration of 12 weeks or less.

Klingner et al. (2006) broadened the inclusion criteria by including studies concentrating on a K–12 population and by not requiring a certain design feature. However, they narrowed down the studies to those that targeted ELLs with LD or ELLs who were struggling readers or “at risk” or “low achieving”, which resulted in a total of seven intervention studies. In contrast, in the current review, two studies targeting ELLs with LD or RD, and 18 studies conducted for struggling readers, were included.

In short, because of different inclusion criteria and 20 more studies published since Cheung and Slavin's (2005) and Klingner et al.'s (2006) reviews, the present review identifies a larger pool of reading instruction for ELLs in preschool to 6th grades than previous reviews; 26 studies with 44 independent samples, which aggregated 18 different instructional practices. Although this number is still insufficient considering the large population of ELLs, it is clear that the literature on reading instruction for ELLs is rapidly growing.

What Reading Instructional Programs are Available for ELLs?

An important goal of this study was to identify what reading instructional programs are available for ELLs. Based on theories of second language development or specifically ELLs' English development (Krashen, 1987, 1988; Cummins, 1981), there are essential instructional practices that should be in place at each developmental stage. Specifically, phonemic awareness and phonics instruction are fundamental at emergent (preschool to mid-first grade) and beginning (kindergarten to early third grades) stages. In addition, building vocabulary should also be a part of early reading instruction at these stages. (Bear et al., 2004; 2007).

Findings of the present study demonstrate that at both tier 1 and tier 2, there are more than 10 instructional programs available that address phonemic awareness and phonics instruction at beginning grades (preschool to 2nd grades). Yet, in spite of

the instructional need for phonemic awareness and phonics components for many struggling readers at upper grades (The National Center for Education Statistics, 2008), there is no phonemic awareness and phonics instruction that has been empirically validated for these grades at either tier 1 or tier 2. In particular, when considering the unique situation of U.S. classrooms in which many ELLs are just learning the English language, and are possibly just learning to read any language, these students may need strong supports in reading as emergent readers. As I indicated in *the stages of ELLs' reading development in English* section in the literature review (Chapter II), when ELLs have reading difficulties, their teachers should evaluate the student's ELL language development stage first, rather than counting on his/her grade or age. For example, if a 5th grade age-level Hmong student immigrated to the U.S. recently, never received formal schooling in his home country, and is beginning to learn to read in English, his teacher should consider his stage emergent, and plan instruction accordingly. If this instruction is not sufficient to help the student improve, moving to a more intensive tier of intervention might be considered.

Another obvious lack of research is on vocabulary instruction. Although the importance of vocabulary knowledge has been emphasized (e.g., Helman, 2008) , there are a few vocabulary instructional programs available for ELLs at tier 1; two programs for preschool to 2nd graders and three for students in upper grades. Based on the stages of English language acquisition development, explicit vocabulary

instruction is required to build English vocabulary even at the emergent stage, which is preschool, kindergarten and mid-first grade (Bear et al., 2004; 2007).

Last, because this review was limited to group studies, studies with single-subject design (n=16) were not included. Yet, these were the only studies that addressed the kinds of intensive, individualized interventions that would be classified as tier 3. As a result, this review cannot provide any information, including the availability and the effect, about tier-3 instruction for which studies most often adopt single-subject designs. However, because of its resemblance to traditional special education with the form of individualized intensive instruction, further meta-analytic review is needed for studies with single-subject designs.

Do Reading Instructional Practices Work for ELLs?

The overall effect of reading instruction for ELLs synthesized from studies comparing ELLs in intervention to those in control groups were moderate, with a mean ES of 0.50. This means that the mean score of the intervention group is at the 69th percentile rank of the control group. The 95% confidence interval was 0.37 to 0.63. Thus, the findings of the review demonstrate that ELLs clearly benefit from reading instructional programs examined in studies included in this review, compared to receiving traditional or regular reading instruction.

Moreover, the overall effect of reading instruction for ELLs synthesized from studies comparing ELLs and L1 students who received the same intervention was not significantly different from zero, with a mean ES of 0.07. This suggests that the reading programs included in Data Set 2 have no differential effect for ELLs and L1 students. Yet, this encouraging finding is based only on 9 studies with 5 instructional programs; thus, this somewhat positive interpretation should be viewed with caution.

Because this review was limited to studies published in peer-reviewed journals, the publication bias must also be addressed. The most critical question might be: are these positive effects upwardly biased due to publication bias? As explained in the introduction section, the publication bias (Torgerson, 2006) or file drawer problem (Rosentahl, 1979; Howell, 2008) represents a tendency to publish only those studies with statistically significant results; negative or near neutral results are almost never published. The potential effect of publication bias in research synthesis is that effect sizes may be inflated (e.g., Shadish & Baldwin, 2005). The publication bias is closely related to the apples-and-oranges criticism, which means that, by including poorly designed studies in a research synthesis, the summaries of the review become meaningless. On the one hand, when a researcher includes almost all studies in the synthesis regardless of their quality, which has been strongly suggested by Glass (2000), the mixing of apples and oranges would render the simplified summaries meaningless. On the other hand, when a researcher

includes only high quality studies in the synthesis, which has been strongly advocated by Slavin (1986, 1987), the publication bias could render an inflated overall effect.

To deal with this dilemma, I chose an empirical approach; including studies published in peer-reviewed journals but examining the relation between research quality and effect size. Still, there might be a selection bias, because I possibly left out the lowest quality or the lowest effect studies associated with the extreme ESs, and thus it might cause a restricted range for calculating the coefficient. However, this selection bias would result in downward bias due to its restricted range. The publication bias usually refers to upward bias to the average effect size by including only published studies with significant effects (Lipsey & Wilson, 2001).

With regard to this relation, Browder et al. (2006) found that not all studies with high quality in their review had strong effects and vice versa, and emphasized the importance of evaluating the quality of the research design in conjunction with effect size. Moreover, both study quality and effect size are the criteria of evidence-based practice proposed by Gersten et al. (2005), which was the major topic of this study.

The findings of this study revealed that the correlation between quality and effect size was not statistically significant although the direction is negative as expected, r = -0.331 (p=.052). This is consistent with Graham and Perin's (2007) findings in their synthesis on writing intervention. They applied quality indicators

proposed by Gersten et al. and found no relation with $r = .07$. No relation between study quality and effect size implies that two criteria of evidence-based practice are not overlapping or redundant, but that both are necessary in their own right. Furthermore, the weak but negative direction of relation also implies that researcher's efforts to maintain the rigor of research design by controlling extraneous variables may lead to lower effects (Swanson & Hoskyn, 1998; Simmerman & Swanson, 2001). Swanson & Hoskyn found that studies failing to report psychometric information on participants with learning disabilities yielded significantly higher effect sizes than those studies reporting that information. Similarly, Simmerman and Swanson reported that the presence of desirable methodological features in a study significantly corresponds with lower effect sizes. The present review supports their findings; When considering no relation with negative direction between quality and effect, and the relatively high quality of studies included, it can be concluded with more confidence that the overall positive effect for ELLs found in this review was not inflated by publication bias.

What Factors Do or Do Not Affect the Strength of Effects?

This review identified several moderators that strengthen the effect sizes; (1) English as a language of intervention, (2) 45 min or less for session length, (3) upper grade level, (4) single grade, (5) average or higher SES, (6) phonemic awareness

outcome measure, and (7) lower quality of study. Each moderator and its implication is discussed below.

Language of Intervention

With the other predictors held constant, the ES was significantly higher for studies in which the intervention language was English, compared to studies in which the intervention language was Spanish. The weighted ES for interventions in English was 0.46 (30 samples, 144 ESs), and that of Spanish was 0.18 (5 samples, 31 ESs). One study using both Spanish and English was categorized as Spanish intervention.

Yet, the low effect of Spanish intervention should be interpreted with caution because the instructional goals of Spanish intervention in two samples (Vaughn & Linan-Thomson et al., 2006; Vaughn & Cirino et al., 2006) were to improve Spanish, not English. English outcome measures were added to examine cross-language transfer. Indeed, the weighted ES of three Spanish interventions targeting English improvement (ES=0.98; Avila & Sadoski, 1996; Bernhard et al., 2006; Zhang & Schumm, 2000) was much higher than that of two Spanish interventions targeting Spanish improvement (ES=.16). It was also much higher than that of overall English intervention (ES=0.46). Thus, the low effect of Spanish intervention seems to be caused by the two studies in which Spanish interventions were implemented mainly to improve Spanish. However, because of the small

number of studies examining the impact of Spanish as language of intervention, firm conclusions cannot be drawn.

Length of Session

The effect of intervention was significantly larger for studies in which a session lasted 45 min or less, compared to studies in which sessions lasted over 45 min. The weighted ES for sessions lasting 45 min or less was 0.52 (27 samples, 93 ESs), and that of over 45 min was 0.21 (8 samples, 82 ESs). When considering the almost even number of ESs included in this analysis, these results are conclusive. Thus, it is assumed that when a session lasts more than 45 minutes, students in preschool to sixth grades might lose attention or teachers' instructional focus might diffuse, which might cause lower effect.

Grade Level and Multiple Grades Included

ELL students' grade variable was also a significant predictor of effect size. With the other predictors held constant, the effect was significantly higher when ELL participants in 3rd to 6th grade were included in the studies, compared to preschool to 2nd grades included, and when a single grade was included in the studies, compared to multiple grades. The weighted ES for 3rd to 6th graders was 0.58 (11 samples, 27 ESs) was significantly higher than preschool to 2nd grade, 0.33 (24 samples, 148 ESs). The weighted ES for studies including only a single grade was 0.36 (23

samples, 135 ESs), which was significantly higher than multiple grades included, 0.29 (12 samples, 40 ESs). The stronger effect of including only one grade is plausible due to possible homogeneity of instructional components and its resulting instructional intensity. Thus, this single grade effect might interact with the other variables, in particular those at the instructional level. However, in general, it is hard to determine the reason for these grade effects.

SES Level

ELL students' SES level was a significant predictor of effect size. With the other predictors held constant, the effect of intervention was significantly higher when ELL participants' SES level was average or above, compared to low SES. The weighted ES for average or higher SES was 0.46 (16 samples, 52ESs), and that of low SES was 0.31(19 samples, 123ESs). The SES level has broad connotations for educational environment and resources (Evans, 2004). The relative contribution of poverty and racial minority on the achievement gap has been one of the hottest issues for educators and policy-makers; indeed, the "achievement gap" has been discussed as a matter of SES and race. Across the U.S., a gap in academic achievement persists between students from ethnic minority and low SES backgrounds and their white counterparts (The Economic Policy Institute, 2009). On the one hand, poverty has been shown to be a major contributing factor to low achievement. For example, before even entering kindergarten, the average cognitive

score of children in the highest SES group are 60% above the scores of the lowest SES group (Lee & Burkam, 2002). An income increase of $1,000 has been found to associate with an increase in reading test scores of 3.6% of a standard deviation (Dahl & Lochner, 2005). On the other hand, some researchers stated that poverty does not completely explain why the performance of African American and Latino students lags behind, insisting there are not just SES achievement gaps, but racial achievement gaps; students in lower SES schools, the majority of whom are Hispanic (73%) and African American (10%), showed substantially lower academic performance than those in higher SES schools having relatively smaller proportion of these ethnic groups (15% and 4%) (e. g., California Department of Education, 2007). However, this present review shows there is strong SES effect with studies conducted only with a relatively homogeneous ethnic minority group; the majority of ELLs in the review were Hispanic (73%).

Construct of Outcome Measures

The effect size was different by the type of outcome construct. The weighted ESs were 0.41 for phonemic awareness (n=26), 0.33 for phonics (n=72), 0.38 for fluency (n=27), 0.34 for vocabulary (n=11), and 0.32 for comprehension measures (n=39); compared to 0.41 for phonemic awareness, the other components have lower ESs. The stronger effect on phonemic awareness measures in ELL population may

reflect that ELL students' actual growth on this measure is larger than the others and/or the sensitivity of the measure to the growth.

However, before interpreting in that way, it is important to examine possibilities of interactions with other significant predictors. Although five outcome constructs did not differ in language of instruction and session length variables, they differ in beginner grade and single–grade studies. Specifically, significantly more students assessed with phonemic awareness measures were in beginner grades (which has significantly lower ESs than upper grades) than those on the vocabulary and comprehension measures ($p < .01$); thus, this interaction cannot have an impact on higher ESs in this measure. Moreover, students assessed with this measure were less frequently in single grade status (which has significantly higher ES than multiple grades) than other students assessed with other measures; thus this interaction also cannot account for the higher ES in this measure. Only significant differences that phonemic awareness has, compared to other outcome measures, were its frequency on the alignment with instructional components (significantly more phonemic awareness measures were aligned with instruction than vocabulary measures were), and on the use of standardized score variable (significantly more phonemic awareness measures were reported as raw scores than comprehension measures were). Overall, it does not appear that the other significant predictors affect the strong effect on the phonemic awareness measure. Thus, the plausible reason of

the higher effect on this measure is that ELL students show larger growth on phonemic awareness and/or the measure has greater sensitivity to students' growth.

Quality of study

With the other predictors held constant, the mean ESs decreased significantly with the increase of the quality of studies (coefficient =-0.18). Along with the marginally significant correlation between study quality and effect size in a negative direction (r = -0.331), the significance of study quality variable means that researchers' efforts to maintain the rigor of research design by controlling extraneous variables are highly likely to lower effects.

Variables That did Not Affect the Strength of Effect

Several variables did not moderate the effect size: properties of outcome measures (standard or raw scores; aligned with instructional components or not); intervention tier (tier 1 vs. tier 2), number of instructional components included (1 to 5), ELL students' home language (Spanish vs. not Spanish), ELL students' reading performance (low vs. at average or above), timing of outcome measurement (immediate vs. follow-up). -On all these variables, reading instruction was equally effective for both or all conditions with the overall mean ES of 0.50. Among those, no effect for number of instructional components is surprising (1, ES=0.66; 2, ES=0.54; 3, ES=0.16; 4, ES=0.69; 5, ES=0.28) when considering general

recommendation of providing a complex programs teaching several of these components simultaneously for ELLs (e.g., August & Shanahan, 2006).

Which Instructional Programs Can be Considered Evidence-Based?

One of the primary purposes of this review was to identify evidence-based reading instruction for ELLs in elementary grades. I identified instruction as evidence-based using the criteria proposed by Gersten et al. (2005): instruction can be considered evidence based if it is supported by at least two high quality studies and a weighted effect size significantly greater than zero. According to these criteria, the Keyword method and Proactive Reading in English are considered evidence-based, and Peer-Assisted Learning Strategies (PALS) is promising. Each instructional approach is described in more detail below, followed by a brief discussion of issues related to establishing an evidence base of instructional programs.

Keyword method. The Keyword method was supported by two quality studies and a weighted ES of 1.07 (Avila & Sadoski, 1996; Zhang & Schumm, 2000). These studies examined effects of the keyword method in comparison to the rehearsal method on the vocabulary learning of 5th grade ELLs in general classrooms. Interestingly, the language used in two of three independent samples was Spanish to acquire English vocabulary.

The strength of the Keyword method is in its nature of simplicity and explicitness as tier 1 instruction in general classroom (Bear et al., 2007). Although the interventionists, which were the authors, had to make the vocabulary booklets, they were able to train ELL students with ease. The booklets consisted of English words to be learned, English keywords, the English definitions, and pictures to associate the keyword with the meaning of the word to be learned for two training sessions, which lasted 30 min and 60 min respectively. Students believed that they found out secrets for learning new words. Keyword method has been shown to be effective for learning foreign languages, including English, but its effect on long-term memory of English vocabulary has been questioned (e.g., McDaniel, Pressely, Dunay, 1987; Pressley & Levin, 1981; Pressley, Levin, Hall, Miller, Berry, 1980; Pressley, Levin, Kuiper, Bryant, & Michener, 1982, Wang & Thomas, 1995). Moreover, the authors of keyword studies ((Avila & Sadoski, 1996; Zhang & Schumm, 2000) also questioned long-term vocabulary retention, because, students' vocabulary learning was evaluated at immediate and one-week time intervals. In addition, the outcome measures did not have evidence of reliability and validity.

Another evidence-based reading program identified in this synthesis is Proactive Reading, researched by Vaughn and her colleagues (Vaughn & Linan-Thomson et al., 2006; Vaughn & Mathes et al, 2006; Vaughn & Cirino et al., 2006). Of the four samples they reported, two used Spanish and two used English as language of intervention. The English intervention reported in two quality studies

was evidence-based with the weighted ES of 0.35, but Spanish intervention was not with the ES of 0.04. Yet, the Spanish intervention was designed to improve students' Spanish achievement, not English performance; the English outcome was evaluated only to examine cross-language transfer.

Proactive reading was a tier 2 intervention for small groups of 3 to 5 struggling readers implemented daily (50 min) for 7 months. The strength of the instruction resides in its systematic and explicit nature and its comprehensiveness including all of the key reading instructional components except vocabulary. To achieve this comprehensiveness, intervention teachers received hours of professional development and coaching prior to and during the implementation, which may reduce its feasibility.

Peer-Assisted Learning Strategies. Peer-Assisted Learning Strategies (PALS) was identified as promising practice. Although the weighted ES was moderate with 0.57 from four high quality studies, the *Z*-score was not significant because of a large standard error (0.33), which results in a 95% confidence interval that includes zero. The large standard error came from two samples in which ELL group composition was different from other studies in this review; a small portion of L1 students included (37% for Calhoon et al., 2006; 21% for Calhoon et al., 2007) and a low percentage of LEP students (28%) among ELL participants (Calhoon et al, 2007). When removing these samples from its aggregation, the weighted ES increases to as 0.40 (SE=0.09, Z=4.64, p<0.01) and the 95% confidence interval is

0.06 to 0.74, which meets the evidence-based criterion. The advantages of PALS are its wide application and comprehensiveness; it has been researched with kindergarten to 6th grade, shown its effects as both tier 1 and 2 approaches, is explicit, skill-based instruction. However, like Proactive Reading, PALS implementation requires professional development and support, which may reduce its feasibility.

Establishing an evidence base. With the aggregation of studies conducted for the same instructional program, three programs discussed above were identified as evidence-based or promising practices. The research pool for the aggregation was 29 studies with 45 comparisons or samples with 225 outcome measures or effect sizes. In a recent special issue of *Exceptional Children* (Vol. 75, No. 3, 2009), 5 reviews that focused on the topic of evidence-based practices for reading, math, writing, and behavior were conducted. Each of the authors concluded that meeting the criteria of evidence-based practices is not easy. The number of evidence-based practices identified in this review illustrates how difficult it is to acquire this title; three programs were identified as evidence-based or promising practices out of 13 programs in 29 studies.

In light of these findings and those of other reviewers, some might insist upon the need to reduce the rigor of evidence-based criteria. In fact, some researchers have criticized the WWC's (2008) standards as overly rigorous, resulting in few practices with positive effects identified, and have referred to the WWC as the "nothing works clearinghouse," (e.g., Viadero & Huff, 2006, p. 8). However,

other researchers argue for maintaining high standards; as one reviewer stated: "rather than reducing the criteria for quality, our research community should carefully consider how to reinforce efforts to improve the quality of research" (Chard et al., 2009, p.278). Still, the quality indicators appear to need some refinement, such as operational definitions for each criterion to be a reliable measurement tool, weighting the components according to their importance, and guidelines for how to deal with multiple components for each indicator item (Cook et al., 2009).

It is important to note that the criteria used in this review are "proposed" or recommended to use and follow, but not the absolute standard of choice for ELLs' reading instruction. As Cook et al. cautioned (p. 380), "evidence-based standards should interface with the professional wisdom of teachers to maximize the outcomes of students." In particular, in the field of special education, *individualized approach* is as always as critical as *standardized method*.

How Do these Findings Differ from those of the "What Works Clearinghouse" ?

Recently, the What Works Clearinghouse (WWC, 2009) conducted and released a review on reading instruction for ELLs along with the ratings of effectiveness, which is almost identical to what the present review pursues. The WWC was established in 2002 by the U.S. Department of Education's Institute of

Education Sciences to provide educators, policymakers, researchers, and the public with a source of scientific evidence about "what works" in general education. Besides instruction for ELLs, WWC reviewed a variety of topics, such as adult literacy, beginning reading, elementary school math, middle school math, dropout prevention, and charter education. Because WWC is viewed as an educational authority, it is critical to discuss the similarities and differences between the two reviews.

For the topic of ELLs, WWC identified 73 studies of 33 programs that qualified for the review by using following inclusion criteria; (1) only randomized controlled trials and quasi-experimental studies (i.e., quasi-experiments with equating, regression discontinuity designs, and single subject research), (2) interventions and instructional strategies focusing on ELL students in grades K-6, and (3) outcome domains are reading achievement, mathematics achievement, and English language development. Of 33 programs, WWC offered the intervention reports for 12 programs that met their inclusion criteria, including 10 programs on reading and 2 programs on English language development. The outcome measures on English language development were excluded in the present review; the 10 programs on reading achievement evaluated by both reviews are summarized in Table 16 and compared in Table 17.

As shown in Table 16, three programs in the WWC report were not included in my review: (1) the Bilingual Cooperative Integrated Reading and Composition

(BCIRC) program conducted by Calderon (1998) in which the outcome measure was not classified by any of the key components of reading instruction, (2) the Success for All program evaluated by Chambers (2004) but not published in a journal, and (3) the Vocabulary Program conducted by Carlo et al. (2004) in which no data to compute ES (mean, *SD*, *n*) were reported.

Table 17 presents a comparison between the WWC review and the present review in terms of the evaluation for study quality and effectiveness. For study quality, the WWC classifies studies as meeting evidence standards or meeting evidence standards with reservations. Only randomized controlled studies can meet evidence standards without reservation. Quasi-experimental studies with equating, regression discontinuity designs, and single subject research and randomized controlled studies with methodological limitations can meet evidence standards with reservations. In contrast, according to Gersten et al.'s (2005) criteria for study quality, at least 9 out of 10 quality scores could be considered high quality. Four programs with 8 of quality score in my review were rated as "evidence-standard with reservation"; and 1 program having 7 of quality score in my review was rated as "evidence standard without reservation" in the WWC report. Thus, the criteria for quality in my review are a bit more rigorous than that of the WWC.

Table 16

Summaries of Studies Included in WWC and My Reviews

		WWC Review		My Review	
	Program	no of study	authors	no of study	reason of discrepancy
1	BCIRC	1	Calderon1998	0	excluded because of outcome measure
2	Enhanced Proactive Reading	2	(1) Vaughn,Cirino et al.,2006 (2) Vaughn,Mathes et al., 2006	4	2 more studies in Spanish intervention included
3	Fast Forword	2	(1)Scientific Learning Corportion2004 (2) Troia2004	1	(1) not published in peer-reviewed journal
4	Instructional Conversation & Literature Log	2	(1) Saunders & Goldberg1999 (2) Saunders1999	1	(2) excluded because of outcome measure
5	PALS	1	Saenz et al. 2005	4	3 more studies included
6	Read Natually	1	Denton 2004	1	same
7	Read Well	1	Denton 2005	1	same
8	ReadingMastery/Correcti veReading	1	Gunn2000	1	
9	Success for All	1	Chambers 2004	0	not published in a journal
10	VocabularyProgram	1	Carlo et al. 2004	0	excluded because of no data for ES reported

Table 17

Evaluation Comparison for Study Quality and Effectiveness in WWC and My Reviews

	Quality			Effectiveness		EBP
	WWC Review		My Review	WWC Review	My Review	
Program	Design	Evidence Standard(ES)	quality	Rating of Effectiveness	Weigted ES	My Review
Enhanced proactive Reading	(1)&(2) randomized control trails	ES without/ with Reservation	10,10 (10,10)	Potentially Positive*	0.50	*EBP*
Fast Forword	(1) RCT (2) quasi	ES without/ with Reservation	8	No discernible	0.06	Not meet
Conversation & Literature Log	(1) RCT (2) quasi	ES with Reservation	8	Potentially Positive*	0.32	Not meet
PALS	RCT	ES	9 (10,10,10)	Potentially Positive*	0.49(0.57)	*promising*
Read Natually	quasi	ES with Reservati	8	No discernible	0.00	Not meet
Read Well	RCT with differential attrition	ES with Reservation	8	Potentially Positive *	0.26	Not meet
ReadingMastery/CorrectiveReading	RCT	ES	7,8	Potentially Positive *	0.41	Not meet

Note. RCT = Randomized Control Trial, * Potentially Positive = not significanat, but ES>.25;

For effectiveness, the WWC uses five categories: statistically significant positive effects, potentially positive effects, indeterminate effects, potentially negative effects, and statistically significant negative effects. "Potentially" means that effects are educationally meaningful although not statistically significant; specifically, an effect size of greater than ±0.25. The criteria proposed by Gersten et al. consider the weighted effect size significantly greater than 0 for an evidence-based practice, and a 20% confidence interval for the weighted effect size greater than zero for promising practice, along with at least two high quality studies supporting the program. By comparison, all the effect sizes greater than 0.25 in my review were rated as potentially positive in WWC, and two ES near 0 were rated as not discernible. Thus, ratings of effectiveness from two reviews are congruent. However, the criteria proposed by Gersten et al. requires both standards of high quality studies *and* effect size; only two programs were identified as evidence-based, and one was rated as promising in my review. Thus, the evidence-based practice criteria proposed by Gersten et al. seems to be much more stringent than that of WWC. Therefore, considering that some researchers have criticized the WWC's (2008) standards as overly rigorous resulting in few practices with positive effects identified, and further referred to the WWC as the "nothing works clearinghouse," (e.g., Schoenfeld, 2006; Viadero & Huff, 2006, p. 8), the criteria proposed by Gersten et al. might be called as "nothing ever works" criteria.

However**,** the number of practices identified by the WWC and the present review is not as negative as we expect from the criticism that called the WWC as the

“nothing works clearinghouse". In terms of study quality, 16 (22%) out of 73 studies qualified for the WWC”s review have met the *evidence standards*, with 8 *without reservation* and 8 *with reservation*. In terms of effectiveness, among 12 reading programs for which the intervention reports were available, 5 programs (42%) were rated as *potentially positive* based on reading achievement outcome measures, and 4 programs (33%) were *potentially positive* on English language development outcome measures. In my review, 17 (49%) out of 35 studies met the criteria of high quality, and 3 (23%) out of 13 programs were identified *evidence-based or promising practice*.

Therefore, I strongly agree with the idea that rather than reducing the rigor of the quality criteria, researchers should carefully consider how to reinforce efforts to improve the quality of research (Chard et al., 2009, p.278). Yet, the resulting issue is that we might have a rather small inventory of evidence-based practices to recommend to be used by classroom teachers. This is a bit troublesome and problematic. Thus, the research community might need to brainstorm how to revise the criteria proposed by Gersten et al. while at the same time reinforcing efforts to improve the quality of research.

Conclusion

This meta-analytic review systematically and rigorously examined the availability of reading instruction for ELLs in preschool to sixth grades; determined the

effectiveness of that instruction; and identified the evidence-based practices. The review was limited to experimental and quasi-experimental studies published in peer-reviewed journals.

The greatest contribution of meta-analysis is to draw important insights from what might otherwise be confusing in the literature by handling a large number of studies, which might be overwhelming in traditional reviews (Banda & Therrien, 2008; Bangert-Drowns 2004 Lipsey & Wilson, 2001; Walker, Hernandez, & Kattan, 2008). This review demonstrates that overall, ELLs clearly benefit from reading intervention, with the moderate effect size of 0.50 based on 13 programs from 35 studies with 175 ESs.

Moreover, meta-analysis is capable of finding relationships across studies that are obscured in other approaches. In this review, several factors that affect the impact of reading instruction were identified; English as a language of intervention, 45 min or less for session length, upper grade level, single grade, and average or higher SES. Because the WWC report did not look at contextual factors (i.e., predictors) as this review did, it seems that identifying moderators that affect the strength of effect of this study is a particularly valuable contribution.

In addition, a new understanding gained from this review is the relation between study quality and effect size; there was marginally significant correlation in a negative direction, r = -0.331. In a similar vein, the study quality was a significant predictor of effect size in a negative direction. Thus, this relation implies that researchers' efforts to

maintain the rigor of research design by controlling extraneous variables are highly likely to lower effects. Moreover, it implies that two criteria of evidence-based practice are not overlapping or redundant, but that both are necessary in their own right.

Last, but not least important, this study identified three evidence-based or promising practices from 13 programs; Keyword method, Proactive Reading and Peer-Assisted Learning Strategies. This is certainly not sufficient for classroom teachers of ELLs. Thus, the research community should continue to reinforce efforts to improve the quality of research while seeking strong support for rigorous research from policy makers to empower ELL students who are otherwise likely to be at-risk readers by providing them with evidence-based reading instruction.

REFERENCES

References marked with an asterisk indicate studies included in this meta-analysis

*Almaguer, I. (2005). Effects of dyad reading instruction on the reading achievement of hispanic third-grade english language learners. *Bilingual Research Journal, 29*(3), 509-526.

August, D & Hakuta, K (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academics Press.

August, D., & Shanahan, T. (Eds.). (2006) *Developing literacy in second-language learners: Report of the National Literacy Panel on language minority children and youth.* Mahwah, NJ: Lawrence Erlbaum Associate.

*Avila, E., & Sadoski, M. (1996). Exploring new application of the keyword method to acquire English vocabulary. *Language Learning, 46*, 379-395.

Banda, D. R. & Therrien, W. J. (2008). A teachers' guide to meta-analysis. *Teaching Exceptional Children, 41*, 66-71.

Baker, K., & de Kanter, A. (1981). *Effectiveness of bilingual education: A review of the literature.* (Final draft report). Washington, DC: Office of Technical and Analytic Systems, U.S. Department of Education.

Bear, D. R. (1992). The prosody of oral reading and stages of word knowledge. In S. Templeton & D. Bear (Eds.). *Development of orthographic knowledge and the foundation of literacy: A memorial Festschrift for Edmund H. Henderson*. Hillsdale, NJ: Lawrence Erlbaum.

Bear, D. R., Helman, L. A., Templeton, S., Invernizzi, M. & Johnston, F. (2007). *Words their way with English learners: Word study for phonics, vocabulary and spelling instruction.* Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Bear, D.R., Invernizzi, M., Templeton, S., & Johnston, F.R. (2004). *Words their way: Word study for phonics, vocabulary, and spelling instruction* (3rd Ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Beretvas, S. N., & Paster, D.A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*, 75-95.

*Bernhard, J. K., Cummins, J., Campoy, F. I., Ada, A. F., Winsler, A., & Bleiker, C. (2006). Identity texts and literacy development among preschool English language learners: Enhancing learning opportunities for children at risk for learning disabilities. *Teachers College Record, 108*, 2380-2405.

Browder, D. M., Wakeman, S. Y., Spooner, F., Ahlgrim-Delzell, L., & Algozzine, B. (2006). Research on reading instruction for individuals with significant cognitive disabilities. *Exceptional Children, 72*, 392–408.

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

*Calhoon, M. B., Al Otaiba, S., Cihak, D., King, A., & Avalos, A. (2007). Effects of a peer-mediated program on reading skill acquisition for two-way bilingual first-grade classrooms. *Learning Disability Quarterly, 30*(3), 169-184.

*Calhoon, M. B., Al Otaiba, S., Greenberg, D., King, A., & Avalos, A. (2006). Improving reading skills in predominantly Hispanic Title 1 first-grade classrooms: The promise of Peer-Assisted Learning Strategies. *Learning Disabilities Research & Practice, 21,* 261–272

Carlo, M.S., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D. N., Lively. T J., & White. C. E. (2004). Closing the gap: Addressing the vocabulary needs of English language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*, 188-215.

Case, R.E., & Taylor. S. S. (2005). Language difference or learning disability? Answers from a linguistic perspective. *Clearing House, 78,* 127-130.

Center for Research on Education, Diversity, and Excellence. (2004). *Research evidence: Five standards for effective pedagogy and student outcomes.* (Technical Report No. G1)> Santa Cruz: University of California.

Chall, J. S. (1983). *Stages of reading development*. New York: McGraw-Hill.

Cheung, A., & Slavin, R. E. (2005). Effective reading programs for English language learners and other language-minority students. *Bilingual Research Journal, 29*(2), 241-267.

*Chiappe. P., Siegel. L.S., & Wade-Woolley. L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading, 6*, 369–400

Clay, M. M. (1987). Learning to be learning disabled. New *Zealand Journal of Educational Studies, 22*, 155-173.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly, 23,* 509-531.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children, 75*, 365-383.

Council for Exceptional Children (2007) *Position on Response to Intervention (RTI): The unique role of special education and special educators*. . Retrieved May 21, 2009, from http://www.cec.sped.org/AM/Template.cfm?Section=Home&CONTENTID=9237&TEMPLATE=/CM/ContentDisplay.cfm

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California Department of Education (Ed.), *School and language minority students: A theoretical framework* (pp. 3-50). Los Angeles, CA: Evaluation, Dissemination, and Assessment Center.

Dahl, G., & Lochner, L. (2005, August). *The impact of family income on child*

achievement(Discussion Paper No. 1305-05). Madison, WI: Institute for Research on Poverty. Retrieved May 21, 2009, from http://www.irp.wisc.edu/publications/dps/pdfs/dp130505.pdf

DeCoster (2004). DeCoster, J. (2004). *Meta-analysis Notes.* Retrieved March 3, 2009, from http://www.stat-help.com/notes.html

*Denton, C. A., Anthony, J. L., Parker, R., & Hasbrouck, J. E. (2004). Effects of two tutoring programs on the English reading development of Spanish-English bilingual students. *The Elementary School Journal, 104*, 289-305.

Economic Policy Institute. (2009). *Efforts to Close Achievement Gap Must Begin Much Earlier, New Report Shows*. Retrieved May 21, 2009, from http://www.epi.org/pages/press_releases_inequality_gate093002/

Ehri, L. C. (1997). Learning to read and learning to spell are one and the same, almost. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to spell: Research, theory, and practice across languages* (pp. 237-269). Mahwah, NJ: Lawrence Erlbaum Associates.

*Ehri, L. C., Dreyer, L. G., Flugman, B., & Gross, A. (2007). Reading rescue: An effective tutoring intervention model for language minority students who are struggling readers in first grade. *American Educational Research Journal, 44*, 414-448.

Evans, G. W. (2004, February/March). The environment of childhood poverty. *American Psychologist. 59*, 77-92

Eysenck, H. J. (1994). Systematic reviews: Meta-analysis and its problems. *British Medical Journal, 309*, 789-792.

Fender, M. (2003). English word recognition and word integration skills of native Arabic and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics, 24*, 289–315.

Fillmore, L. W., & Snow, C. E. (2002). What teachers need to know about language. In C. T. Adger, C. E. Snow, & D. Christian (Eds.), *What teachers need to know about language* (pp. 7–53). Washington, DC: Center for Applied Linguistics and Delta Systems Co., Inc

Fry, R. (2007). *How far behind in math and reading are English language learners?* (Report June 6, 2007). The Pew Hispanic Center, Washington, D.C.

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August and T. Shanahan (Eds.). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language Minority Children and Youth* (pp.153-174). Mahwah, NJ: Lawrence Erlbaum Associate.

Genesee, F., Lindholm-Leary, K., Saunders, W. & Christian, D. (2005). English language learners in U. S. schools: An overview of research findings. *Journal of Education for Students Placed at Risk, 10,* 363-385.

*Gerber, M., Jimenez, T., Leafstedt, J., Villaruz, J., Richards, C., & English, J. (2004). English reading effects of small-group intensive intervention in Spanish for K-l

English learners. *Learning Disabilities Research & Practice, 19*, 239-251.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality Indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71,* 149-164.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V (2000). *Meta-analysis at 25*. Retrieved May 21, 2009, from http://glass.ed.asu.edu/gene/papers/meta25.html

Glass G. V., McGaw, B., Smith M. L. (1981). *Meta-analysis in social research*. Newbury Park: Sage Publications

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In Cooper, H.,& Hedges, L. V. (Eds.). The Handbook of Research Synthesis. The RussellSage Foundation.

Graham, S. & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Education Psychology, 99*, 445-476.

Greene, J. P. (1997). A meat-analysis of the Rossell and Baker review of bilingual education research. *Bilingual Research Journal, 21*, 103-122.

*Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic student in early elementary school. *Journal of Special Education, 34*, 90-103.

Harwell, M. (1997). An empirical study of Hedges' Homogeneity Test. *Psychological Methods*,*2*, 219-231.

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-anlayses and some remedies. *Journal of Experimental Education*, 76, 403-428.

Hedges, L.V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

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

Helman, L. A. (2004). Building on the sound system of spanish: Insights from the alphabetic spellings of english-language learners. *The Reading Teacher, 57*(5), 452-460.

Helman, L. A. (2005). Using literacy assessment results to improve teaching for english-language learners. *The Reading Teacher, 58*, 668-677.

Helman, L. A., & Burns, M. K. (2008). What does oral language have to do with it? Helping young english-language learners acquire a sight word vocabulary. *The Reading Teacher, 62*, 14-19.

Helman,L. A.(2008). English word needed: Creating research-based vocabulary instruction for English learners. In A.E. Farstrup, & S.J. (Eds.). *What research has to say about vocabulary instruction*. Newark, DE: International reading Association.

Hillocks, G. (1986). *Research on written composition: New directions for teaching.* Urbana, IL: National Council of Teachers of English.

Horner, R. H., Carr, E.G., Halle, J., & Odom, S. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165-179.

Howell, R. T. (2008). The File drawer problem in reliability generalization *Educational and Psychological Measurement, 68*, 120-128

Hunter J. E., Schmidt F. L., Jackson G. B. (1982). *Meta-analysis: cumulating research findings across studies.* Beverly Hills , Ca: Sage Publications.

Individuals with Disabilities Education Improvement Act, P. L. 108-446 U.S.C. (2004).

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate Linear Model for Meta-Analysis. Psychological Methods, 1(3), 227-235

*Kamps, D., Abbott, M., Greenwood, C., Arreaga-Mayer, C., Wills, H., Longstaff, J. et al., (2007). Use of evidence-based, small-group reading instruction for English language learners in elementary grades: Secondary-tier intervention. *Learning Disability Quarterly, 30,* 153-168.

Klingner, J. K., Artiles, A. J., & Barletta, L. M. (2006). English language learners who struggle with reading, language acquisition or LD? *Journal of Learning Disabilities, 39*, 108-128.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge, England: Cambridge University Press

Krashen, S. (1987). *Principles and practice in second language acquisition.* New York: Prentice Hall International.

Krashen, S. (1988). *Second language acquisition and seond language learning.* New York: Prentice HIall International.

Krashen, S., & Brown, C. L. (2005). The ameliorating effects of high socioeconomic status: A secondary analysis. *Bilingual Research Journal, 29*(1), 185-196.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school.* Washington, D.C.: Economic Policy Institute, 2002.

Landis, J.R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in *Biometrics*. Vol. 33, pp. 159--174

Lane, K. L., Kalberg, J. R., & Shepcaro, J. C. (2009). An examination of the evidence-base for function-based intervention for students with emotional and/or behavioral disorders attending middle and high schools. *Exceptional Children, 75*, 321-342.

*Leafstedt, J. M., Richards, C. R., & Gerber, M. M. (2004). Effectiveness of explicit phonological-awareness instruction for at-risk English learners. *Learning Disabilities Research and Practice, 19*, 252-261

*LeSaux, N. K. & Siegel, L. S. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology, 39*, 1005-1019.

Lipsey, M. W., & Wilson, D. B (2001). Practical meta-analysis: Applied social research methods series. Thousand Oaks, CA: Sage Publications, Inc.

McCardle, P., Mele-McCarthy, J., Cutting, L., Leos, K. (2005). Englsh language

learners and learning disabilities: Research agenda and implications for practice. *Learning Disabilities Research & Practice, 20*, 68-78.

McCardle, P., Mele-McCarthy, J., Cutting, L., Leos, K., & D'Emilio, T. (2005). Learning disabilities in English language learners: Identifying the issues. *Learning Disabilities Research & Practice, 20*, 1-5.

McDaniel, M.A., Pressely, M., Dunay, P. K. (1987). Long-Term Retention of Vocabulary After Keyword and Context Learning. *Journal of Educational Psychology, 79*, 87-89.

McGaw, B. (1997). Meta-analysis. In J. P. Keeves (Ed.), *Educational research, methodology,and measurement : an international handbook (2nd ed.)* Cambridge: Cambridge University Press

*McMaster, K. L., Kung. S-H., Han. I., & Cao. M. (2008). Peer-Assisted Learning Strategies: A "Tier 1" approach to promoting English learners' response to intervention. *Exceptional Children, 74,* 194-214.

Minnesota Department of Education (2003). *The ELL companion to reducing bias in special education evaluation*. Retrieved January 15, 2008, from http://education.state.mn.us/MDE/Learning_Support/Special_Education/Evaluation_Program_Planning_Supports/Cultural_Linguistic_Diversity/ELL_Companion_Manual/index.html

National Clearinghouse for English Language Acquisition (2006). *The growing numbers of LEP students: 2001-2002.* Retrieved January 15, 2008, from http://www.ncela.gwu.edu/stats/4_toplanguages /languages.html

National Institute of Child Health and Human Development. (2000).

National Literacy Panel on Language Minority Children and Youth. (2006).

National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read.* Bethesda, MD: National Institute of Child Health and Human Development.

No Child Left Behind Act, Public Law No. 107-110, 115 Stat. 1425, 2002 U.S.C. (2002).

Northwest Regional Educational Laboratory. (2003). *Overview of second language acquisition theory*. Retrieved May 20, 2009, from http://www.nwrel.org/request/2003may/overview.html

Pressley, M., Levin, J. R., Hall, J. W., Miller, G. E., & Berry, J. K. (1980). The keyword method and foreign word acquisition. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 163-173.

Pressley, M., & Levin, J. R. (1981). The keyword method and recall of vocabulary words from definitions. *Journal of Experimental Psychology: Human, Learning and Memory, 7*, 72-76.

Pressley, M., Levin, J. R., Kuiper, R. A., Bryant, S. L., & Michener, S. (1982). Mnemonic versus non-mnemonic vocabulary-learning strategies: additional comparisons. *Journal of Educational Psychology, 74*, 693-707.

*Proctor, C. P., Dalton, B., & Grisham, D. L. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with

embedded strategy instruction and vocabulary support. *Journal of Literacy Research, 39*, 71-93.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA:Sage.

Reed, B. & Railsback, J. (2003). *Strategies and resources for mainstream teachers of English language learners.* Retrieved May 20, 2009 from http://www.nwrel.org/request/2003may/textonly.html

Rolstad, K., Mahoney, K., & Glass, G. V. (2005). The big picture: A meta-analysis of program effectiveness research on English Language Learners. *Educational Policy, 19*, 572-594

Rosenthal, R. (1979), "The file drawer problem and tolerance for null results", *Psychological Bulletin, 86,* 638–641

Rosenthal, D. A., Hoyt, W. T., Ferrin, J. M., Miller, S., & Cohen, N. D. (2006). Advanced methods in meta-analytic research: Applications and implications for rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 49*, 234-246.

Rossell, C. H., & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English, 30*, 7–69.

Rudner, L., Glass, G.V, Evartt, D.L., & Emery, P.J. (2002).*A user's guide to the meta-analysis of research studies.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

*Sanez, L. M., Fuchs. L. S., & Fuchs, D. (2005). Peer-Assisted Learning Strategies for English language learners with learning disabilities. *Exceptional Children, 71*, 231-247.

*Saunders, W. M., & Goldenberg, C. N. (1999). Effects of instructional conversations and literature logs on limited- and fluent-English-proficient students' story comprehension and thematic understanding. *The Elementary School Journal, 99*, 277-301.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13–21.

Shadish, W. R. & Baldwin, S. A. (2005). Effects of behavioral marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 73*, 6-14.

*Silverman, R. D. (2007). Vocabulary development of english-language and english-only learners in kindergarten. *The Elementary School Journal, 107*(4), 365-383.

Simmerman, S., & Swanson, H. L. (2001). Treatment outcomes for students with learning disabilities: How important are internal and external validity? *Journal of Learning Disabilities, 34*(3), 221-236.

Slavin, R. E. (1986). *Ability Grouping and Student Achievement in Elementary Schools" A Best-Evidence Synthesis.* (Rep. No. 1). Baltimore, Md.: Johns Hopkins University, Center for Research on Elementary and Middle Schools.

Slavin, R. E. (1987). "Mastery Learning Reconsidered." *Review of Educational Research* 57, 2: 175-213.

Slavin, R. E. & Cheung, A. (2005). A Synthesis of Research on Language of Reading Instruction for English Language Learners. *Review of Educational Research, 75*, 247–284.

Slavin, R. E. (2008). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37*(1), 47-50.

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research, 75*, 247-284.

*Stuart, M. (1999). Getting ready for reading: Early phoneme awareness and phonics teaching improves reading and spelling in inner-city second language learners. *The British Journal of Educational Psychology, 69*(pt4), 587-605.

*Stuart, M. (2004). Getting ready for reading: A follow-up study of inner city second language learners at the end of key stage 1. *The British Journal of Educational Psychology, 74*(pt1), 15-36.

Suri, H. (2000). A critique of contemporary method of research synthesis, *Post-Script, 1*, 49-55.

Suri, H., & Clarke, D. (2009). Advancements in research synthesis methods: From a methodologically inclusive perspective. *Review of Educational Research, 79*(1), 395-430.

Swanson, H. L., & Hoskyn M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 68*, 277-321.

Swanson, H. L., & Jerman, O. (2006). Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research, 76*, 249-274.

Swanson, H. L., Zheng, X. & Jerman, O. (2009). Working memeory, short-term memory, and reading disabilities: A selective meta-analysis of the literature. *Journal of Learning Disabilities, 42*, 260-287.

Task Force on Evidence-Based Interventions in School Psychology. (2003). *Procedural and coding manual for review of evidence-based interventions.* Division 16 of the American Psychological Association.

Test, D. W., Fowler, C. H., Brewer, D. M., & Wood, W. M. (2005). A content and methodological review of self-advocacy intervention studies. *Exceptional Children, 72*, 101-125.

Torgerson, C. J. (2006). Publication bias: The achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*(1), 89-102.

*Troia, G. A. (2004). Migrant students with limited English proficiency: Can Fast ForWord Language™ make a difference in their language skills and academic achievement? *Remedial and Special Education, 25*, 353-366.

U.S. Census Bureau (2009). *S1401. School enrollment*. Retrieved May 20, 2009, from http://factfinder.census.gov/servlet/STTable?_bm=y&-geo_id=01000US&-

qr_name=ACS_2007_3YR_G00_S1401&-ds_name=ACS_2007_3YR_G00_

U.S. Census Bureau (2009). *S1601. Language spoken at home*. Retrieved May 20, 2009, from http://factfinder.census.gov/servlet/STTable?_bm=y&-geo_id=01000US&-qr_name=ACS_2007_3YR_G00_S1601&-ds_name=ACS_2007_3YR_G00_&-_lang=en&-redoLog=false&-CONTEXT=st

U.S. Department of Education, & National Institute of Child Health and Human Development. (2003). *National symposium on learning disabilities in English language learners. Symposium summary.* Washington,DC: Authors.

U.S. Department of Education, National Center for Education Statistics. (2006). *The condition of education 2006*. NCES 2006-071. Washington D.C.: U.S. Government Printing Office. Retrieved January 15, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006071

*Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., & Hagan, E. C., et al. (2006). Effectiveness of a Spanish intervention and an English intervention for English-language learners at risk for reading problems. *American Educational Research Journal, 43*, 449-487.

*Vaughn, S., Linan-Thompson, S., Mathes, P.G., Cirino, P.T., Carlson, C.D., & Pollard-Durodola, S.D. et al. (2006). Effectiveness of Spanish intervention for first-grade English language learners at risk for reading difficulties. *Journal of Learning Disabilities, 39*, 56-73.

*Vaughn, S., Mathes, P., Linan-Thompson, S., Cirino, P.T., Carlson, C.D., Pollard-Durodola, S. D., et al. (2006). Effectiveness of an English intervention for first-grade English language learners at risk for reading problems. *The Elementary School Journal, 107*, 153-180.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S., Chen, R., Pratt, A., Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*, 601-638.

Viadero, D., & Huff, D. J. (2006). "One stop" research shop seen as slow to yield views that educators can use. *Education Week, 26*(5), 8–9.

Wang, A.Y. & Thomas, M.H. (1995). Effect of keywords on long-term retention: help or hindrance? *Journal of Educational Psychology, 87*, 468-475.

What Works Clearinghouse. (2009). *English language learners*. Retrieved May 20, 2009, from http://ies.ed.gov/ncee/wwc/reports/Topic.aspx?tid=10

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55*, 269–317.

Wortman, P. M., & Bryant, F. B. (1985). School desegregation and black achievement: An integrative review. *Sociological Methods and Research, 13*, 289-324.

Yang, J. (2006). A meta-analysis of the effect of interventions to increase reading fluency among elementary school students. Unpublished doctoral dissertation.

Vanderbilt University, Nashville, TN.

*Zhang, Z., & Schumm, S. S. (2000). Exploring effects of the keyword method on limited English proficient students' vocabulary recall and comprehension. *Reading Research and Instruction, 39*, 202-221.

APPENDICES

Appendix A

Quality Indicators for Group Studies Proposed by Gersten et al., (2005)

Quality Indicators for Describing Participants

1. Was sufficient information provided to determine/confirm whether the participants demonstrated the disability(ies) or difficulties presented?
2. Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?
3. Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

Quality Indicators for Implementation of the Intervention and Description of Comparison Conditions

4. Was the intervention clearly described and specified?
5. Was the fidelity of implementation described and assessed?
6. Was the nature of services provided in comparison conditions described?

Quality Indicators fir Outcome Measures

7. Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance?
8. Were outcomes for capturing the intervention's effect measured at the appropriate times?

Quality Indicators for Data Analysis

9. Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?
10. Did the research report include not only inferential statistics but also effect size calculations?

Appendix B

Key Instructional Ccomponents (National Reading Panel, 2000): Related Instructional Activities and Outcome Measures Used in the Study

1. ***Phonemic awareness***: Phonemic awareness refers to the ability to focus on and manipulate the phonemes in spoken words. (Phonemic awareness is the ability to isolate and manipulate the individual sounds in words). Phonemes are the smallest units making up spoken language. English consists of about 41 phonemes. Phonemes combine to form syllables and words. A few words have only one phoneme, such as a (a) or oh (o). Most words consist of a blend of phonemes, such as go (g-o) with two phonemes, check (ch-e-ck) with three phonemes, or stop with four phonemes (s-t-o-p).

    - **Instructional (intervention) activities for *Phonemic Awareness*** include:
        - Phonemic onset
        - Rhyming (for example, pictures of "pan, fan", ask "which one rhyme with pan and fan, strawberry or ran?) =recite rhymes
        - Segmenting (for example, teacher "soap:, student "s-oa-p")
        - Blending (for example, teacher "s-oa-p:, student "soap")

    - **Outcome measures for *Phonemic Awareness*** include:
        - Blending (for example, teacher "s-oa-p:, student "soap")
        - Segmenting (for example, teacher "soap:, student "s-oa-p"), including Phoneme Segmentation Fluency (PSF) in DIBELS
        - The Comprehensive Test of Phonological Processes (CTOPP; Torgesen & Wagner), which includes elision (phoneme deletion, for example, "Say cowboy without boy"), blending, and segmenting
        - The *Lindamood Auditory Conceptualization Test* (LAC; Lindamood& Lindamood, 1979)

2. ***Phonics***: Phonics instruction is a way of teaching reading that stresses learning how letters correspond to sounds and how to use this knowledge in reading and spelling. (Phonics instruction is to teach children learn the relationships between the letters of written language and the sound of spoken language) (Letter-sound relations, symbol-sound relations, decoding, word-attack skills, and alphabetic knowledge (print knowledge)). Phonics instruction can be provided systematically. Systematic phonics instruction occurs when children receive explicit, systematic instruction in a set of pre-specified associations between letters and sounds. Children are taught how to use these associations to read, typically in texts containing controlled vocabulary.

- **Instructional (Intervention) activities for *Phonics*** include:
  - (read out) Sight words
  - (read out) Dolch words
  - Fry lists
  - Prefix
  - Suffix
  - High frequency words
  - Chunking
  - Decoding
  - Retaining (sight words)
  - Repeated exposure to words (sight words)
  - Letter-sound relations (Letter-Naming, Naming)

- **Outcome measures for *Phonics*** include:
  - Rapid Letter Naming(RLN), Rapid Letter Sound (RLS) in PALS studies, Rapid Automatized Naming (RAN) test (Denckla & Rudel, 1974)
  - Nonsense Word Fluency (NWF) in DIBELS (confirmed by KM, 11/25), pseudo-word repetition
  - Letter Naming and Sound Identification in Proactive Reading
  - Word Identification and Word attack in Woodcock Reading Mastery Test-Revised(WRMT-R)
  - Letter Naming Fluency (LNF) in DIBELS
  - The *Woodcock-Mufioz Language Survey, letter- word identification subtest*
  - RES test in ReadingRescue
  - LAP-D in EarlyAuthorProgram
  - The reading subtest of the Wide Range Achievement Test–3 (WRAT–3;Wilkinson, 1995); word and letter recognition
  - ***Not include Writing measures (Spelling & Dictation test)***

3. Fluency: Fluency is the ability to read a text accurately and quickly. Reading fluency is one of several critical factors necessary for reading comprehension, but is often neglected in the classroom. If children read out loud with speed, accuracy, and proper expression, they are more likely to comprehend and remember the material than if they read with difficulty and in an inefficient way. Two instructional approaches have typically been used to teach reading fluency. One, guided repeated oral reading, encourages students to read passages out loud with systematic and explicit guidance and feedback from their teacher. The other, independent silent reading, encourages students to read silently on their own, inside and outside the classroom, with little guidance or feedback from their teachers.

- **Instructional (Intervention) activities for *Fluency*** include:
    - Read sentences, passages, or stories (not just separate words) orally
    - Read at home
    - Just plain practice reading; repeated reading
    - Errors (not referring to the measure)
    - Errors in accuracy
    - Corrective Reading
    - Pencil tap (refers to accuracy)

- **Outcome measures for *Fluency*** include:
    - Read sentences, passages, or stories (not just separate syllables or words) orally
    - Oral Reading Fluency (ORF) in DIBELS, which read words in connected text
    - The Comprehensive Reading Assessment Battery (CRAB), word correct score, in PALS

4. ***Vocabulary***: Children learn the meanings of words indirectly through daily engagement in oral language, listening to adults read to them, and reading on their own. Vocabulary can be developed when students are explicitly taught both individual words and word learning strategies

    - **Instructional (Intervention) activities for *Fluency*** include:
        - Keyword method

    - **Outcome measures for *Vocabulary*** include:
        - Peabody Voc test
        - The Oral Expression subtest of the *Oral and Written Language Scales* (OWLS; Carrow-Woolfolk, 1995)
        - *Picture Vocabulary subtest* of the *Woodcock-Mufioz Language Surve,*
        - *Picture Vocabulary subtest* of the WLPB-R (Woodcock, 1991)

5. ***Text Comprehension*:** Reading comprehension is very important to the development of children's reading skills and therefore to their ability to obtain an education. In carrying out its study of reading comprehension, the NRP noted three main themes in the research on the development of reading comprehension skills. First, reading comprehension is a complex cognitive process that cannot be understood without a clear description of the role that vocabulary development and vocabulary instruction play in the understanding of what has been read. Second,

comprehension is an active process that requires an intentional and thoughtful interaction between the reader and the text (text comprehension instruction). Third, the preparation of teachers to better equip students to develop and apply reading comprehension strategies to enhance understanding is intimately linked to students' achievement in this area.

- **Instructional (Intervention) activities for *Comprehension*** include***:***
    - Meta-cognitive (comprehension) strategies, which includes reviewing, sequencing, summarizing, stating main ideas, or predicting.
    - Question answering
    - Question generation
    - Summarization (=shrinking)
    - Recall of ideas = Retell stories
    - Predicting (& checking) outcomes
    - Context clues
    - Experience that relates to prior knowledge
    - Literature logs
    - Instructional conversation

- **Outcome measures for *Comprehension*** include:
    - The Comprehensive Reading Assessment Battery (CRAB), question correct score and maze correct score, in PALS
    - The Listening Comprehension subtest of the *Oral and Written Language Scales* (OWLS; Carrow-Woolfolk, 1995) (Q. is LC a comprehension or phonics measure or either by context?; similarly, is Auditory comprehension scale, which is PLS-R in Bernhard2006, a phonics or comp?)
    - Essays (Explanation and exemplification essays)
    - The Passage comprehension subtest of WRMT-R
    - Verbal analogies
    - PLS-R in earlyAuthorProgram
    - ITBS, reading test
    - La Prueba Spnish Reading Test
    - Reading Strategies
    - Sentence completion

Appendix C

Study Level Coding Manual

clm	Variable Nan	Descriptions & Codes
A	ID	Study ID Number. Assign a unique identification number to each study. If a report presents two independent studies, i.e., two independent outcome studies with different subjects, then add a decimal to the study ID number to distinguish each study within
B	author	Last name of the first author
C	pub_type	What type of publication is the report? If two separate reports are being used to code a single study, code the type of the more formally published report (i.e., book or journal article).
		1 peer-reviewed journal
		2 unpublished dissertation
		3 book or book chapter
		4 other (specify)
D	pub_yr	What is the publication year? If two separate reports are being used to code a single study, code the publication year of the more
	Instuctional component descriptors	
E	int_title	Write the name of intervention.
F	int_language	Language of intervention
		1 English
		2 Spanish
		3 English+Spanish
		9 other (specify)
G	int_act	instructional activities (see "5 instructional component)
H	int_PA	Phonemic Awareness (Refer to the operational definition by National Reading Panel)
		1 Yes
		0 No
I	int_ph	Phonics (operational definition by National Reading Panel)
		1 Yes
		0 No
J	int_flu	Fluency (operational definition by National Reading Panel)
		1 Yes
		0 No
K	int_voc	Vocabulary (operational definition by National Reading Panel)

		1	Yes
		0	No
L	int_comp		Comprehension
		1	Yes
		0	No
M	int_tier		Tier (Preventative level)
		1	Tier 1 (primary prevention; classwide instruction for all students ingeneral classroom)
		2	Tier 2 (Secondary prevention; supplemental instruction for at-risk students or nonresponders to Tier 1 instruction)
		3	Tier 3 (Tertiary prevention; invidualized instruction for nonresponders to tier 2 instruction or students with disabilities)
	ELL participa		(Note: For study with ELLs & L1s, L1 students will be coded at Study Descriptors section)
O	age_mean		Mean age of ELL participants in total . Specify the mean age the beginning of intervention. (digit, XX.XX)
P	gr_total		Grades of participant in total. Specify all the grades, coding only the lowest and the highest. For example, if grades are kindergraten to grade 3rd, code "K3"; grade 2 to grade 5, code "25"
Q	gr_Single		Single grade or mixed grades
		1	Single grade
		2	Mixed grades
R	gr_K		Kindergarten
		1	Yes
		0	No
S	gr_1		Grade 1
		1	Yes
		0	No
T	gr_2		Grade 2
		1	Yes
		0	No
U	gr_3		Grade 3
		1	Yes
		0	No
V	gr_4		Grade 4
		1	Yes
		0	No
W	gr_5		Grade 5

		1	Yes
		0	No
X	gr_6		Grade 6
		1	Yes
		0	No
Y	N_total		write the total number of ELL participants used for analysis, to calculate the percentage of FEMALE ELL participants ,
Z	N_female		write the number of female ELLs used for analysis . If not specified, leave blank
AA	ratio_female		Leave blank, will be automatically calculated (digit, XX.XX)
AB	ELL_lang		Home language of ELLs
		1	All Spanish (i.e., include Spanish + L1; Spanish dominant LEP)
		2	Mixed ELL (i.e., Spanish & non Spanish, just report as Mixed ELL)
		3	NoInfo (just report as ELL)
		9	Other, specified (specify in next column)or just report as non-Spanis
AC	ELL_langSp		When above coding is 2 or 9, specify all other home languages besides Spanish (e.g., Korean, Chinese)
AD	ELL_ability		ELL participants' performance level in reading
		1	General Education, low
		2	General Education, full range (low to high)
		3	General Education, at-risk readers
		4	Special Education (LD, RD)
		5	not specified;NoInfo
		6	Middle
		7	high
AE	ELL_SES		ELL participants' SES level
		1	low
		2	middle
		3	high
		4	whole range
		5	not specified;NoInfo
		9	the other
AF	ELL_prfi		ELL participants' English language proficiency level
		1	low (defined by LEP, or score, or both)
		2	whole range
		4	not specified;NoInfo
		9	the other

Research Design Descriptors

AG	sch_set		The location of schools
		1	urban, metropolitan
		2	rural
		3	suburban
		4	not specified;NoInfo
AH	assign_type		Unit of assignment to conditions; Select the code that describe the unit of assignment to intervention and control conditions
		1	random assignment (after matching, stratification, blocking) by individual
		2	random assignment (after matching, stratification, blocking) by classroom/teacher
		3	matched pair in existing random-assigned groups
		4	(matched/non-matched) exisiting groups (i.e., quasi-ex. design)
		5	n/a, no assignment (i.e., single subject design)
		6	not specified;NoInfo
AI	design		Design of the study, Recode from "assign_type" in SPSS
		1	experimental group (recode from 1, 2, 3 in "assign_type")
		2	quasi_experimental (recode from 4 in "assign_type")
		3	single-subject (recode from 5 in "assign_type")
AJ	equi_pre		Is there any evidence of group equivalence at pre-test? i.e., random assignment a/o statistics that show equivalence
		1	Yes
		2	No
AK	diff_pre		Pretest differences between groups if statistics were repored.
		1	no differences at all
		2	some differences, which means significant differences on less than half of variables (tests, demographics)tested.
		3	big differences, which means significant differences on more than half of variables(tests, demographics)tested.
AL	N_tot_begin		Total sample size at the beginning of study
AM	N_tot_end		Total sample size at the end of study that used for analysis
AN	att_rate		Autmatically calculated from AL & AM
AO	att_why		reason of attrition
		1	systematic drop-out
		2	non-systematic/random drop-out
		3	not specified;NoInfo
AP	N_cont		control group sample size at the end of study
AQ	N_exp		intervention, experimental group sample size at the end of study

Intervention Descriptors in detail (WHEN no informaation, leave blank for open-end

AR	leng_Min		length of each session in minutes; when reported as range, min
AS	leng_Max		length of each session in minutes; when reported as range, max
AT	sess_freq		avr. frequency of sessions per week (ex., 3-4/wk=3.5/wk)
AU	dur_int		whole duration of intervention in weeks (ex., 1yr=9Mo=6wks)
AV	sess_total		automatically calculated by AT*AU, or when total frequency is reported without information about AT & AU
Aw	grouping		grouping format
		1	one to one tutor (use one computer)
		2	pair (2)
		3	small group (3 or more)
		4	whole class
AX	who_int		Select the code that describe who implement the intervetion
		1	teacher (classroom teachers, resourceroom teacher)
		2	researcher (authors, including faculty a/o graduate students)
		3	undergr-/graduate students who are not authors
		4	others from outside (i.e., hired & trained teachers)
		5	not specified;NoInfo
AY	tra_int		training or workshop for the interventionist
		1	Yes, described
		2	not specified;NoInfo
AZ	loc_int		Location that the intervention takes place
		1	classroom
		2	resource room
		3	student's home
		4	not specified;NoInfo
BA	asst_int		whether consultation/assistance is available
		1	Yes; described
		2	not specified;NoInfo
BB	fid_yes		whether fidelity is evaluated
		1	Yes; described
		2	not specified;NoInfo
BC	fid_freq		how many times fidelity was evaluated
BD	fid_score		Mean of fidelity score, percentage (digit, XX.XX)
BE	fid_interr		whether interater agreement for fidelity is reported
		1	Yes; reported
		2	not specified;NoInfo
BF	interr_score		Mean of interater agreement score, if reported, in percentage (digit, XX.XX)

BG	control_des		whether the authors describe what the instruction in control looks like
		1	Yes; described
		2	not specified;NoInfo

Appendix D

Studies Excluded in This Meta-Analysis

1. The study employed single-subject design (n = 16).

Al Otaiba (2005) How effective is code-based reading. tutoring in English for English learners and preservice teacher-tutors? *Remedial and Special Education, 26*, 245-254

Bliss, S. L., Skinner, C. H., & Adams, R. (2006). Enhancing an english language learning fifth-grade student's sight-word reading with a time-delay taped-words intervention. *The School Psychology Review, 35*(4), 663-670.

de la Colina, M. G., Parker, R. I., Hasbrouck, J. E., & Lara-Alecio, R. (2001) Intensive intervention in reading fluency for at-risk beginning Spanish readers. *Bilingual Research Journal*, *25*, 503-538.

Gilbertson. D., & Bluck, J. (2006) Improving responsiveness to intervention for English-language learners: A comparison of instructional pace on letter naming rates. *Journal of Behavioral Education,* 15:131–147.

Gibertson. D., Maxwell, J., & Hughes, J. (2007) Evaluating responsiveness to intervention for English-Language learners: A Comparison of response modes on letter naming rates. *Journal of Behavioral Education, 16,* :259–279.

Greenwood. C. R., Arreaga-Mayer. C., Utley. C.A., Gavin. K. M., & Terry. B.J. (2001). Classwide peer tutoring learning management system: Applications with elementary-level English language learners. *Remedial and Special Education, 22*, 34-47.

Haager, D. & Windmueller, M. P. (2001). Early reading intervention for English language learners at-risk for learning disabilities: Student and teacher outcomes in an urban school. *Learning Disability Quarterly, 24*, 235-250.

Hapstak, J., & Tracy, D. H. (2007). Effects of assisted-repeated reading on students of varying reading ability: A single-subject experimental research study. *Reading Horizons Journal, , 47,* 315-334.

Jitendra, A. K., Edwards, L. L., Starosta, K., Sacks, G., Jacobson, L. A., & Choutka, C. M. (2004). Early reading instruction for children with reading difficulties: Meeting the needs of diverse learners. *Journal of Learning Disabilities, 37,* 421-439.

Li, D., & Nes, S. (2001). Using paired reading to help ESL students become fluent and accurate readers. Using paired reading to help ESL students become fluent and accurate readers. *Reading Improvement, 38,* 50-61.

Malloy, K. J., Gilbertson, D., & Maxfield, J. (2007) Use of brief experimental analysis for selecting reading interventions for English language learners. *School Psychology Review, 36,* 291-310

O'Donell, P., Weber, K. P., & McLaughlin, T. F. (2002) Improving Correct and Error Rate and Reading Comprehension Using Key Words and Previewing: A Case Report with a Language Minority Student, *Education and Treatment of Children,*

26, 237-254.
Rousseau, M ., & Tam, B. K. Y. (1991). The efficacy of previewing and discussion of key words on the oral reading proficiency of bilingual Learners with speech and language impairment. *Education & Treatment of Children, 14*, 199-210.
Rousseau, M ., & Tam, B. K. Y. (1993). Increasing reading proficiency of language-minority students with speech and language impairments. *Education & Treatment of Children, 16,* 254-262.
Santoro, L. E., Jitendra, A. K., & Starosta, K. (2006). Reading well with "read well": Enhancing the reading performance of english language learners. *Remedial and Special Education, 27*(2), 105-115.
Tam, K. Y., Heward, W. L., & Heng, M. A. (2006). A reading instruction intervention program for English language learners who are struggling readers, *The Journal of Special Education, 40,* 79-93.

2. The study did not have a control group although it employed group experimental design (n =4).

Neal, J. C., Kelly, P. R., & (1999) The Success of reading recovery for English language learners and Descubriendo La Lectura for bilingual students in California, *Literacy Teaching and Learning, 4(2),* 81-107
Poulsen, R., Hastings, P., & Allbritton, D. (2007). Tutoring bilingual students with an automated reading tutor that listens. *Journal of Educational Computing Research, 36*(2), 191-221.
Silverman, R. (2007). A comparison of three methods of vocabulary instruction during read-alouds in kindergarten. *The Elementary School Journal, 108*(2), 97-113.
Smith, P. E. (1996). Reading recoveruy and children with Englsih as a second language. *Literacy, teaching and Learning, 2*, 61-78..

3. The study did not provide adequate data to compute the standard mean difference ES (n = 3)

Carlo, M. S., August, D., & McLaughlin, B. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.
D'Angiulli, A., Siegel, L. S., & Maggi, S. (2004). Literacy instruction, SES, and word-reading achievement in English-language learners and children with English as a first language: A longitudinal study. *Learning Disabilities Research and Practice*, *19*, 202-213.
Perez (1981). Oral language competence improve reading skills of Mexican American third graders. *Reading teacher, 35*, 24-27.

4. The study did not include reading outcome measures of key components of reading ($n = 2$).

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

Muniz- Swicegood, M. (1994). The effects of metacognitive reading strategy training on the reading performance and student reading analysis strategies of third grade bilingual students. *Bilingual Research Journal, 18*, 83-97.