EFFECTIVE READING PROGRAMS FOR ENGLISH LANGUAGE LEARNERS

A Best-Evidence Synthesis

Robert E. Slavin
Johns Hopkins University

Alan Cheung
Success for All Foundation

Report No. 66

December 2003

This report was published by the Center for Research on the Education of Students Placed At Risk (CRESPAR), a national research and development center supported by a grant (No. R117-D40005) from the Institute of Education Sciences (IES, formerly OERI), U.S. Department of Education. The content or opinions expressed herein do not necessarily reflect the views of the Department of Education or any other agency of the U.S. Government. Reports are available from: Publications Department, CRESPAR/Johns Hopkins University; 3003 N. Charles Street, Suite 200; Baltimore MD 21218. An on-line version of this report is available at our web site: *www.csos.jhu.edu*.

Copyright 2003, The Johns Hopkins University. All rights reserved.

THE CENTER

Every child has the capacity to succeed in school and in life. Yet far too many children fail to meet their potential. Many students, especially those from poor and minority families, are placed at risk by school practices that sort some students into high-quality programs and other students into low-quality education. CRESPAR believes that schools must replace the "sorting paradigm" with a "talent development" model that sets high expectations for all students, and ensures that all students receive a rich and demanding curriculum with appropriate assistance and support.

The mission of the Center for Research on the Education of Students Placed At Risk (CRESPAR) is to conduct the research, development, evaluation, and dissemination needed to transform schooling for students placed at risk. The work of the Center is guided by three central themes—ensuring the success of all students at key development points, building on students' personal and cultural assets, and scaling up effective programs—and conducted through research and development programs in the areas of early and elementary studies; middle and high school studies; school, family, and community partnerships; and systemic supports for school reform, as well as a program of institutional activities.

CRESPAR is organized as a partnership of Johns Hopkins University and Howard University, and is one of twelve national research and development centers supported by a grant (R117-D40005) from the Institute of Education Sciences (IES, formerly OERI) at the U.S. Department of Education. The centers examine a wide range of specific topics in education including early childhood development and education, student learning and achievement, cultural and linguistic diversity, English language learners, reading and literacy, gifted and talented students, improving low achieving schools, innovation in school reform, and state and local education policy. The overall objective of these centers is to conduct education research that will inform policy makers and practitioners about educational practices and outcomes that contribute to successful school performance.

ABSTRACT

This report reviews experimental studies of reading programs for English language learners, focusing both on comparisons of bilingual and English-only programs and on specific, replicable models that have been evaluated with English language learners. The review method is best-evidence synthesis, which uses a systematic literature search, quantification of outcomes as effect sizes, and extensive discussion of individual studies that meet inclusion standards. The review concludes that while the number of high-quality studies is small, existing evidence favors bilingual approaches, especially paired bilingual strategies that teach reading in the native language and English at the same time. Whether taught in their native language or English, English language learners have been found to benefit from instruction in comprehensive reform programs using systematic phonics, one-to-one or small group tutoring programs, cooperative learning programs, and programs emphasizing extensive reading. Research using longitudinal, randomized designs is needed to understand how best to ensure reading success for all English language learners.

ACKNOWLEDGMENTS

We would like to thank Margarita Calderón, Diane August, Tim Shanahan, Isabel Beck, Jay Greene, Nancy Madden, and Bette Chambers for their comments on an earlier draft, and Susan Davis and the Center for Applied Linguistics for assistance with the literature search.

INTRODUCTION

The reading education of English language learners (ELLs) has become one of the most important issues in all of educational policy and practice. As the pace of immigration to the U.S. and other developed countries has accelerated in recent decades, increasing numbers of children in U.S. schools come from homes in which English is not the primary language spoken. These children represent about 20% of all U.S. students (Van Hook & Fix, 2000). While many children of immigrant families succeed in reading, too many do not. In particular, Latino and Caribbean children are disproportionately likely to perform poorly in reading and in school. As No Child Left Behind and other federal and state policies begin to demand success for all subgroups of children, the reading achievement of English language learners is taking on even more importance. Thousands of schools cannot meet their adequate yearly progress goals, for example, unless their English language learners are doing well in reading. More importantly, American society cannot achieve equal opportunity for all if its schools do not succeed with the children of immigrants.

The great majority of non-English speaking immigrants in the U.S. are of Hispanic origin, and this is also one of the fastest growing of all groups. Hispanics have recently surpassed African Americans as the largest minority group in the U.S. Hispanic students as a whole, including English proficient children in the second generation and beyond, score significantly lower in reading than other students. On the National Assessment of Educational Progress (NAEP; Grigg, Daane, Jin, & Campbell, 2003), which excludes children with the lowest levels of English proficiency from testing, only 44% of Latino fourth graders scored at or above the "basic" level, in comparison to 75% of Anglo students. Only 15% of Latino fourth graders scored at "proficient" or better compared to 41% of Anglos.

There is considerable controversy, among both policymakers and researchers, about how best to ensure the reading success of English language learners. Of course, there are many aspects of instruction that are important in the reading success of English language learners, yet one question has dominated all others: What is the appropriate role of the native language in the instruction of English language learners? In the 1970s and 1980s, policies and practice favored bilingual education, in which children were taught partially or entirely in their native language, and then transitioned at some point during the elementary grades to English-only instruction. Such programs are still widespread, but from the 1990s to the present, the political tide has turned against bilingual education, and California, Arizona, Massachusetts, and other states have enacted policies to greatly curtail bilingual education. Recent federal policies are restricting the amount of time children can be taught in their native language. Among researchers, the debate between advocates of bilingual and English-only reading instruction has been fierce, and ideology has often trumped evidence on both sides of the debate (Hakuta, Butler, & Witt, 2000).

While language of instruction is certainly important, it is only one of many aspects of reading instruction that could affect the achievement of English language learners. Many reviewers of the literature on effective reading instruction for English language learners have argued that quality of instruction is as important or more important than language of instruction. Yet what does this mean in practice? How can we ensure quality reading instruction for all children? Is quality instruction fundamentally different for English language learners than it is for other children?

This report reviews research on effective reading instruction for English language learners in an attempt to apply consistent, well-justified standards of evidence to learn about effective reading instruction for these children. We focus equally on language of instruction and on replicable programs intended to improve the reading achievement of English language learners. This review applies a technique called "best-evidence synthesis" (Slavin, 1986), which attempts to use consistent, clear standards to identify unbiased, meaningful information from experimental studies and then discusses each qualifying study, computing effect sizes but also describing the context, design, and findings of each study. Details of this procedure are described later.

The purpose of this review is to examine the evidence on reading programs for English language learners to discover how much of a scientific basis there is for competing claims about effects of various programs. Our purpose is both to inform practitioners and policymakers about the tools they have at hand to help all English language learners learn to read, and to inform researchers about the current state of the evidence on this topic as well as gaps in the knowledge base in need of further scientific investigation.

LANGUAGE OF INSTRUCTION

For many years, the discussion about effective reading programs for English language learners has revolved around the question of the appropriate language of instruction for children who speak languages other than English. Proponents of native language instruction argue that while children are learning to speak English, they should be taught to read in their native language first, to avoid the failure experience that is likely if children are asked to learn both oral English and English reading at the same time. Children are then transitioned to English-only instruction when their English is sufficient to ensure success, usually in third or fourth grade. Alternatively, many programs teach young children to read both in their native language and in English at different times of the day. There is a great deal of evidence that children's reading proficiency in their native language is a strong predictor of their ultimate English reading performance (Garcia, 2000; Reese, Garnier, Gallimore, & Goldenberg, 2000), and that bilingualism itself does not interfere with performance in either language (Yeung, Marsh, & Suliman, 2000). Advocates also argue that without native language instruction, English language learners are likely to lose their native language proficiency, an important resource in its own right. Opponents, on the other hand, argue that native language instruction interferes with or delays English language development, and relegates children who receive such instruction to a second-class, separate status within the school and, ultimately, within society.

Reviews and research on the educational outcomes of native language instruction have reached sharply conflicting conclusions. In a meta-analysis, Willig (1985) concluded that bilingual education was more effective than English-only instruction. Wong-Fillmore and Valadez (1986) came to the same conclusion. Rossell and Baker (1996) came to the opposite conclusion, claiming that most methodologically adequate studies found transitional bilingual education to be no more effective than English-only programs. Greene (1997) re-analyzed the studies cited by Rossell and Baker and reported that many of the studies they cited lacked control groups, mischaracterized the treatments, or had other serious methodological flaws. Among the studies that met an acceptable

standard of methodological adequacy, including all of the studies using random assignment to conditions, Greene found that the evidence favored programs that made significant use of native language instruction. August and Hakuta (1997) concluded that while research generally favored bilingual approaches, the nature of the methods used and the populations to which they were applied were more important than the language of instruction per se. Program quality, they concluded, was the key. For example, carefully designed, structured immersion programs using only English may have good evidence of effectiveness, but this does not justify "sink or swim" (or "submersion") English-only programs. Well-designed transitional bilingual programs may also benefit children, but there is no justification for poorly designed native language programs that merely give English language learners less demanding instruction. More generally, researchers of all ideological persuasions are converging on the conclusion that the nature and quality of instruction provided to English language learners are at least as important as the language of instruction (see for example, Brisk, 1998; Christian & Genesee, 2001; Goldenberg, 1996; Secada et al., 1998). Quantitative research on the outcomes of bilingual education has diminished in recent years, but policy and practice are still being influenced by conflicting interpretations of research on this topic. The following sections systematically examine this evidence to attempt to discover what we can learn from research to guide policies in this controversial arena.

Immersion and Bilingual Programs

When a child enters kindergarten or first grade with limited proficiency in English, the school faces a serious dilemma. How can the child be expected to learn the skills and content taught in the early grades while he or she is learning English? There may be many solutions, but two fundamental categories of solutions have predominated: *Immersion* and *bilingual education.*

Immersion. In immersion strategies, English language learners are expected to learn in English from the beginning, and their native language plays little or no role in daily reading lessons. Formal or informal support is likely to be given to ELLs to help them cope in an all-English classroom. This might or might not include help from a bilingual aide who provides occasional translation or explanation, a separate English-as-a-Second-Language class to help build oral English skills, or use of a careful progression from simplified English to full English as children's skills grow. Teachers of English language learners might use language development strategies, such as total physical response and realia, strategies for giving students concrete objects, and actions to help them internalize new vocabulary. They might simplify their language and teach specific vocabulary likely to be unfamiliar to ELLs (see Calderón, 2001). Immersion may involve placing English language learners immediately in classes containing English monolingual children, or it may involve a separate class of ELLs for some time until children are ready to be mainstreamed. These variations may well have importance in the outcomes of immersion strategies, but their key common feature is the exclusive use of English texts, with instruction overwhelmingly or entirely in English.

Many authors have made distinctions among different forms of immersion. One term often encountered is "submersion," primarily used pejoratively to refer to "sink or swim" strategies in which no special provision is made for the needs of English language learners. This is contrasted with "structured English immersion," which refers to a well-planned, gradual phase-in of English

instruction relying initially on simplification and vocabulary-building strategies. In practice, immersion strategies are rarely pure types, and in studies of bilingual education, immersion strategies are rarely described beyond their designation as the English-only "control group."

Bilingual Education. Bilingual education differs fundamentally from immersion in that it gives English language learners significant amounts of instruction in reading and/or other subjects in their native language. In the U.S., the overwhelming majority of bilingual programs involve Spanish, due to the greater likelihood of a critical mass of students who are Spanish-dominant and to the greater availability of Spanish materials than those for other languages. For example, children may be taught to read entirely in Spanish in kindergarten and first grade and then transitioned to English. Bilingual programs can be "early-exit" models, with transition to English completed in second or third grade, or "late-exit" models, in which children may remain throughout elementary school in native-language instruction to ensure their mastery of reading and content before transition (see Ramirez, Pasta, Yuen, Billings, & Ramey, 1991). Alternatively, "paired bilingual" models teach children to read in both English and their native language at different time periods each day. Within a few years, the native language reading instruction may be discontinued, as children develop the skills to succeed in English. Willig (1985) called this model "alternative immersion," because children are alternatively immersed in native language and English instruction.

Two-way bilingual programs, or dual language, provide reading instruction in the native language (usually Spanish) and in English both to ELLs and to English speakers (Calderón & Minaya-Rowe, 2003; Howard, Sugarman, & Christian, 2003). For the ELLs, a two-way program is like a paired bilingual model, in that they learn to read both in English and in their native language at different times each day.

A special case of bilingual education is programs designed more to preserve or show respect for a given language than to help children who are genuinely struggling with English. For example, Bacon, Kidd, and Seaberg (1982) studied a Cherokee language program used with children who were of Cherokee ancestry but in many cases did not speak the language. Morgan (1971) studied a program in Louisiana for children whose parents often spoke French at home, but generally spoke English themselves. Such "heritage language" programs are included in this review if the outcome variable in the study is an English reading measure. They should be thought of, however, as addressing a different problem from that addressed by bilingual or immersion reading instruction for children who are limited in English proficiency.

Problems of Research on Language of Instruction

Research on the achievement effects of teaching in the child's native language in comparison to teaching in English suffers from a number of inherent problems beyond those typical of other research on educational programs. First, there are problems concerning the ages of the children involved, the length of time they have been taught in their first language, and the length of time they have been taught in English. For example, imagine that a bilingual program teaches Spanish-dominant students primarily in Spanish in grades K-2, and then gradually transitions them to English by fourth grade. If this program is compared to an English immersion program, at what grade level is

it legitimate to assess the children in English? Clearly, a test in second grade is meaningless, as the bilingual children have not been taught to read in English. At the end of third grade, the bilingual students have partially transitioned, but have they had enough time to become fully proficient? For example, Saldate, Mishra, and Medina (1985) studied Spanish-dominant students in bilingual and immersion schools. At the end of second grade, the bilingual students, who had not yet transitioned to English, scored nonsignificantly lower than the immersion group on English reading. One year later, after transition, the bilingual group scored substantially higher than the immersion group in English reading. Some would argue that even the end of fourth grade would be too soon to assess the children fairly in such a comparison, as the bilingual children need a reasonable time period in which to transfer their Spanish reading skills to English (see, for example, Hakuta, Butler, & Witt, 2000).

A related problem has to do with pretesting. Imagine that a study of a K-4 bilingual transition program began in third grade. What pretest would be meaningful? An English pretest would understate the skills of the bilingual students, while a Spanish test would understate the skills of the English immersion students. For example, Valladolid (1991) compared gains from grades 3-5 for children who had been in either bilingual or immersion programs from kindergarten. These children's "pretest" scores are in fact posttests of very different treatments. Yet studies comparing bilingual and immersion programs are typically too brief to have given the students in the bilingual programs enough time to have fully transitioned to English. Alternatively, many studies begin after students have already been in bilingual or immersion treatments for several years.

The studies that do look at four- or five-year participations in bilingual or immersion programs are usually retrospective (i.e., researchers search records for children who have already been through the program). Retrospective studies also have characteristic biases, in that they begin with the children who ended up in one program or another. For example, children who are removed from a given treatment for systematic reasons, such as Spanish-dominant students removed from English immersion because of their low performance there, can greatly bias a retrospective study.

Many inherent problems relate to selection bias. Children end up in transitional bilingual education or English immersion by many processes that could be highly consequential for the outcomes. For example, Spanish-dominant students may be assigned to Spanish or English instruction based on parent preferences. Yet parents who would select English programs are surely different from those who would select Spanish in ways that would matter for outcomes. A parent who selects English may be more or less committed to education, may be less likely to be planning to return to a Spanish-speaking country, or may feel very differently about assimilation. Thomas and Collier (2002) reported extremely low scores for Houston students whose parents refused to have their children placed in either bilingual or English-as-a-Second-Language programs. Are those scores due to relatively positive effects of bilingual and ESL programs, or are there systematic differences between children whose parents refused bilingual or ESL programs and other children? It is impossible to say, as no pretest scores were reported.

Bilingual programs are more likely to exist in schools with very high proportions of English language learners, and this is another potential source of bias. For example, Ramirez, Pasta, Yuen, Billings, and Ramey (1991) found that schools using late-exit bilingual programs had much higher proportions of ELLs than did early-exit bilingual schools, and English immersion schools had the

smallest proportion of ELLs. This means that whatever the language of instruction, children in schools with very high proportions of ELLs are conversing less with native English speakers both in and out of school than might be the case in an integrated school and neighborhood that uses English for all students because its proportion of ELLs is low. Worst of all, individual children may be assigned to native language or English programs because of their perceived or assessed competence. Native language instruction is often seen as an easier, more appropriate placement for ELLs who are struggling to read *in their first language*, while students who are very successful readers in their first language or are felt to have greater potential are put in English-only classes. This selection problem is most vexing at the point of transition, as the most successful students in bilingual programs are transitioned earlier than the least successful children. A study of bilingual vs. immersion involving third or fourth graders may be seriously biased by the fact that the highest-achieving bilingual students may have already been transitioned, so the remaining students are the lowest achievers.

Finally, a source of bias not unique to studies of bilingual education but very important in this literature is the "file drawer" problem, the fact that studies showing no differences are less likely to be published or to otherwise come to light. This is a particular problem in studies with small sample sizes, which are very unlikely to be published if they show no differences. The best antidote to the "file drawer" problem is to search for dissertations and technical reports, which are more likely to present their data regardless of their findings (see Cooper, 1998).

Because of these inherent methodological problems, an adequate study comparing bilingual and immersion approaches would: a) randomly assign a large number of children to be taught in English or their native language; b) pretest them in their native language when they begin to be taught differentially, either in their native language or in English (typically kindergarten); and c) follow them long enough for the latest-transitioning children in the bilingual condition to have completed their transition to English and have been taught long enough in English to make a fair comparison. Unfortunately, only a few, very small studies of this kind have ever been carried out. As a result, the studies that compare bilingual and English-only approaches must be interpreted with great caution.

Review Methods

This section focuses on research comparing immersion and bilingual reading programs applied with English language learners, with measures of English reading as the outcomes. The review uses a quantitative synthesis method called "best-evidence synthesis" (Slavin, 1986). It uses the systematic inclusion criteria and effect size computations typical of meta-analyses (see Cooper, 1998; Cooper & Hedges, 1994), but discusses the findings of critical studies in a form more typical of narrative reviews. This strategy is particularly well suited to the literature on reading programs for English language learners, because this body of literature is too small and too diverse, both substantively and methodologically, to lend itself to formal meta-analysis.

Literature Search Strategy

The literature search benefited from the assistance of the federally commissioned National Literacy Panel on the Development of Literacy Among Language Minority Children and Youth, chaired by Diane August and Timothy Shanahan. This review, however, is independent of the panel's report, and uses different review methods and selection criteria. Research assistants at the Center for Applied Linguistics (CAL) in Washington, D.C. searched ERIC and other databases for all studies involving language minority students, English language learners, and related descriptors. Citations in other reviews and articles were also obtained. From this set, we selected studies that met the criteria described below.

Criteria for Inclusion

The best-evidence synthesis focused on studies that met minimal standards of methodological adequacy and relevance to the purposes of the review. These were as follows:

1. The studies compared children taught reading in bilingual classes to those taught in English immersion classes, as defined above. Studies of alternative reading programs for English language learners that held constant the language of instruction are discussed in a later section of this review.

2. Either random assignment to conditions was used, or pretesting or other matching criteria established the degree of comparability of bilingual and immersion groups before the treatments began. If these matching variables were not identical at pretest, analyses adjusted for pretest differences or data permitting such adjustments were presented. Studies without control groups, such as pre-post comparisons or comparisons to "expected" scores or gains, were excluded. Studies with pretest differences exceeding one standard deviation were excluded, but those with significant pretest differences less than ES = $\pm$ 1.0 were included if they carried out appropriate adjustments.

   There were several studies in which bilingual and immersion programs were already underway before pretesting or matching. For example, Danoff, Coles, McLaughlin, and Reynolds (1978), in a widely cited study, compared one-year reading gains in many schools using bilingual or immersion methods. The treatments began in kindergarten or first grade, but the pretests (and later, posttests) were administered to children in grades 2-6. Because the bilingual children were primarily taught in their native language in K-1, their pretests in second grade would surely have been affected by their treatment condition. Similarly, several studies tested children in upper elementary or secondary grades who had experienced bilingual or immersion programs in earlier years. These were included if premeasures were available from before the programs began, but in most cases such premeasures are not reported (see, for example, Thomas & Collier, 2002; Curiel, Stenning, & Cooper-Stenning, 1980).

3. The subjects were English language learners in elementary or secondary schools in English-speaking countries. Studies that mixed ELLs and English monolingual students in a way that does not allow for separate analyses were excluded (e.g., Skoczylas, 1972). Studies of other so-

cietal languages would also have been included if they were analogous to the situation of English language learners in the U.S. or Canada (e.g., Turkish children learning to read in Dutch in the Netherlands), but no such studies were found that met the other inclusion criteria. Some studies identified samples of Hispanic or other language minority students without documenting the English proficiency of the children. These studies were included with appropriate explanations. Studies of children learning a foreign language were not included. However, Canadian studies of French immersion have been widely discussed, and are, therefore, discussed in a separate section.

4. The dependent variables included quantitative measures of English reading performance, such as standardized tests and informal reading inventories. If treatment-specific measures were used, they were included only if there was evidence that all groups focused equally on the same outcomes. Measures of outcomes related to reading, such as language arts, writing, and spelling, were not included.

5. This section of the review focused on studies with a treatment duration of at least one school year. For the reasons discussed earlier, even one-year studies of transitional bilingual education are insufficient, because students taught in their native language are unlikely to have transitioned to English. Studies even shorter than this do not address the question in a meaningful way.

Limitations

It is important to note that the review methods applied in this best-evidence synthesis have some important limitations. First, in requiring measurable outcomes and control groups, the synthesis excludes case studies and qualitative studies. Many such descriptions exist, and these are valuable in suggesting programs or practices that might be effective. Description alone, however, does not indicate how much children learned in a given program, or what they would have learned had they not experienced that program. Second, it is possible that a program that has no effect on reading achievement measures might nevertheless increase children's interest in reading or reading behaviors outside of school. However, studies rarely measure such outcomes in any systematic or comparative way, so we can only speculate about them. Finally, it is important to note that many of the studies reviewed took place many years ago, and that both social and political contexts, as well as bilingual and immersion programs, have changed, so it cannot be taken for granted that outcomes described here would apply to outcomes of bilingual and immersion programs today.

Computation of Effect Sizes

Effect sizes were computed for each study. In principle, an effect size is the experimental mean minus the control mean divided by the control group's standard deviation. When this information was lacking, however, effect sizes were estimated using pooled standard deviations, exact t's or p values, or other well-established estimation methods (see Cooper, 1998; Cooper & Hedges, 1994). If effect sizes could not be computed in a study that otherwise qualified for inclusion, the findings were reported. No study was excluded solely on the grounds that it did not provide sufficient information for

computation of an effect size. Because of the small numbers of methodologically adequate studies in each category, no attempt was made to quantitatively pool effect sizes.

Previous Quantitative Reviews

The debate about empirical research on language of instruction for English language learners has largely pitted two researchers, Christine Rossell and Keith Baker, against several other reviewers. Rossell and Baker have carried out a series of reviews and critiques arguing that research does not support bilingual education (see Baker & de Kanter, 1981, 1983; Baker, 1987; Rossell, 1990; Rossell & Ross, 1986). The most comprehensive and recent version of their review was published in 1996. In contrast, Willig (1985) carried out a meta-analysis and concluded that research favored bilingual education, after controls were introduced for various study characteristics. Other reviewers using narrative methods have agreed with Willig (e.g., Wong-Fillmore & Valadez, 1986). Baker (1987) and Rossell and Baker (1996) criticized the Willig (1985) review in detail and Willig (1987) responded to the Baker (1987) criticisms.

In a review commissioned by the Tomas Rivera Center, Jay Greene (1997) carefully re-examined the Rossell and Baker (1996) review. While Rossell and Baker used a "vote-counting" method in which they simply counted the numbers of studies that favored bilingual, immersion, or other strategies, Greene (1997) carried out a meta-analysis in which each study produced one or more effect sizes, the proportion of a standard deviation separating bilingual and English programs. Greene categorized only 11 of the 72 studies cited by Rossell and Baker as methodologically adequate, but among these he calculated an effect size of +0.21 favoring bilingual over English-only approaches on English reading measures. Among five studies using random assignment, Greene calculated an effect size of +0.41 on English reading measures.

As part of this review, we attempted to obtain every study reviewed by Rossell and Baker and by Willig, and independently reviewed each one against the consistent set of standards outlined previously. Consistent with Greene, we found that the Rossell and Baker (1996) review accepted far too many of the articles it summarized. Appendix 1 lists all of the reading studies cited by Rossell and Baker according to categories of methodological adequacy outlined in this report, which closely follow Greene's categorization. As is apparent from the Appendix, only a few of the studies met the most minimal of methodological standards, and most violated the inclusion criteria established by Rossell and Baker (1996) themselves. We found, however, that most of the 16 studies cited by Willig also do not meet these minimal standards. These are also noted in Appendix 1. In itself, this does not mean that the overall conclusions of either review are incorrect, but it does mean that the question of effects of language of instruction on reading achievement must be explored with a different set of studies than the ones cited by either Rossell and Baker or Willig. The Rossell and Baker and Willig studies can be categorized as follows (following Greene, 1997):

1. **Methodologically adequate studies of elementary reading.** These are studies that compared English language learners taught to read using bilingual or immersion strategies, with random assignment or well-documented matching on pretests or other important variables.

2. **Methodologically adequate studies of secondary programs.** We put two secondary school studies (Covey, 1973; Kaufman, 1968) in a separate category.

3. **Canadian studies of French immersion.** Several studies (e.g., Lambert & Tucker, 1972; Genesee & Lambert, 1983) evaluated French immersion programs in Canada. However, since they compared immersion to monolingual English instruction or to brief French-as-a-second-language classes, these are not evaluations of bilingual education.

4. **Studies in which the target language was not the societal language.** In addition to Canadian studies of French immersion in non-francophone areas (e.g., Day & Shapson, 1988), Ramos, Aguilar, and Sibayan (1967) studied various strategies for teaching English in the Philippines.

5. **Studies of outcomes other than reading.** A few studies (e.g., Lum, 1971; Legarreta, 1979) assessed only oral language proficiency, not reading.

6. **Studies in which pretesting took place after treatments were underway.** As noted earlier, many studies (e.g., Danoff et al., 1978; Rosier & Holm, 1980; Rossell, 1990; Thomas & Collier, 2002; Valladolid, 1991) compared gains made in bilingual and immersion programs after the programs were well under way. Both Willig and Rossell and Baker included such studies, and Greene (1997) accepted them as "methodologically adequate," but we would argue that they add little to understanding the effects of bilingual education.

7. **Redundant studies.** Rossell and Baker included many studies that were redundant with other studies in their review. For example, one longitudinal study (El Paso, 1987, 1990, 1992) issued three reports on the same experiment, but it was counted as three separate studies. Curiel's 1979 dissertation was published in 1980, yet both reports are counted. It is important to note that all of these duplicate citation studies found results claimed by Rossell and Baker to favor immersion over bilingual education.

8. **No evidence of initial equality.** Several studies either lacked data on initial achievement, before treatments began, or presented data indicating pretest differences in excess of one standard deviation.

9. **No appropriate comparison group.** Many of the studies included by Rossell and Baker had no control group. For example, Burkheimer, Conger, Dunteman, Elliott, and Mowbray (1989) and Gersten (1985) used statistical methods to estimate where children should have been performing and then compared this estimate to their actual performance. Rossell and Baker's own standards required "a comparison group of LEP students of the same ethnicity and similar language background," yet they included many studies that did not have such comparison groups. Further, many studies included by Rossell and Baker lacked any information about the initial comparability of children who experienced bilingual or English-only instruction (e.g., Matthews, 1979). This includes studies that retroactively compared secondary students who had participated in bilingual or English-only programs in elementary schools but failed to obtain measures of early academic ability or performance (e.g., Powers, 1978; Curiel et al., 1980). Other studies compared obviously non-comparable groups. For example, Rossell (1990) compared one-year gains of English language learners in Berkeley, California, who were in Spanish bilingual or immersion programs, yet 48% of the ELLs, all in the immersion programs, were Asian, while all students in

the Spanish bilingual program (32% of the sample) were, of course, Latino. Also, Legarreta (1979) compared Spanish-dominant children in bilingual instruction to mainly English-dominant children taught in English.

10. **Brief studies.** A few studies cited by Rossell and Baker involved treatment durations less than one year. For the reasons discussed earlier, studies of bilingual education lasting only 10 weeks (Layden, 1972) or four months (Balasubramonian, Seelye, & de Weffer, 1973) are clearly not relevant. Also, all but one of these brief studies failed to meet inclusion standards on other criteria as well (e.g., they lacked pretests or had outcomes other than reading).

The Present Review

This review carries out a best-evidence synthesis of studies comparing bilingual and English approaches to reading in the elementary and secondary grades that meet the inclusion criteria outlined above. These include the methodologically adequate studies cited in the Willig (1985), Rossell and Baker (1996), and Greene (1997) reviews, as well as other studies located in an exhaustive search of the literature, as described previously. The characteristics and findings of these studies are summarized in Table 1.

TABLE 1

Language of Instruction: Descriptive Information and Effect Sizes for Qualifying Studies

Study	Intervention description	Design	Duration	N	Grade	Sample Character-istics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Longitudinal Studies Using Random Assignment										
Plante (1976)	Pairing program; reading taught in Spanish and English	Random assignment	2 yrs	55	1-2, 2-3	Spanish-dominant Puerto Rican students in New Haven, CT	Well matched on Spanish oral vocabulary but C>E in English pretest	**English Inter-American Series**		
								2nd grade	+0.62	+0.43
								3rd grade	+0.24	
Huzar (1973)	Pairing program	Random assignment	2 & 3 yrs	160	1-2, 1-3	Disadvantaged Puerto Rican students in Perth Amboy, NJ	Well matched on IQ, SES, and initial achievement	**English Inter-American Series**		
								2nd grade	+0.01	+0.35
								3rd grade	+0.68	
Maldonado (1994)	Integrated bilingual special education	Random assignment	3 yrs	20	2-4, 3-5	Spanish dominant special education students in Houston TX	Well matched on disability, language proficiency, & family background	**English CTBS**	+2.21	+2.21
Longitudinal Studies Using Matching										
Ramirez et al (1991)	Immersion vs Early Exit	Matched control	4 yrs	Students from various schools	K-3	Spanish dominant LEP students	Fairly well matched on SES and home backgrounds.	**English CTBS**		
								3rd grade	Early=Imm	
Saldate et al (1985)	Bilingual instruction	Matched control	3 yrs	38	1-3	Spanish dominant students in Douglas, AZ	Well matched on pretests	**English tests**		
								MAT (2nd grade)	-0.29	+0.59
								WRAT (3rd grade)	+1.47	
Cohen (1975)	Bilingual program	Matched control	2-3 yrs	90	K-1, 1-2, 1-3	Spanish dominant students in Redwood city, CA	Matched on SES and initial language proficiency	**English Inter-American Series**		
								Cohort 1	E=C	
								Cohort 2	E=C	
Maldonado (1977)	Title VII program vs control	Matched control	5 yrs	126	1-5	Spanish dominant students in six elementary school in Corpus Christi, TX	Matched on SES and number of years in schools	**English (SRAAS)**		
								2nd	E=C	
								3rd	E=C	
								4th	E=C	
								5th	E=C	
Alvarez (1975)	Bilingual versus monolingual	Matched control	2 yrs	147	2	Spanish dominant children in two schools in Austin Texas	Matched on SES and initial language proficiency	**California Achievement Tests**		
								English reading vocab	E=C	
								English reading comp	E=C	
	Corpus Christi Bilingual Education Project	Matched control	2 yrs	171	K-1	Spanish dominant students in Corpus Christi, Texas	Matched on English and Spanish pretests	**English Inter-American Series**	+0.45	+0.45

Study	Intervention description	Design	Duration	N	Grade	Sample Character-istics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Campeau et al (1975)	Houston Bilingual Education Project	Matched control	3 yrs	206	K-2	Spanish dominant students in Houston, TX	Matched on language, SES, and academic achievement	**English Inter-American Series**	+1.00	+1.00
	Alice ISD Bilingual Education Project	Matched control	2 yrs	125	K-1	Spanish dominant students in Alice ISD, Texas	Similar on English pretests but E>C on Spanish pretest	**English Inter-American Series**	+1.06	+1.06
One-Year Studies										
Carlisle & Beeman (2000)	Bilingual program	Matched control	1 yr	36	1	Spanish dominant students	Well matched on home language & SES	**English Woodcock**	+0.07	+0.07
Campeau et al (1975)	Kingsville Bilingual Education Project	Matched control	1 yr	89	K	Spanish dominant students in Kingsville, TX	Matched on SES and ethnic mix	**English Inter-American Series**	E>C	
	Santa Fe Bilingual Education Project	Matched control	1 yr	77	1	Hispanic students in Sante Fe, New Mexico	Pretests, E>C	**English MAT**	+0.28	+0.28
Studies Involving Languages Other Than Spanish										
Morgan (1971)	Bilingual program	Matched control	1 yr	193	1	French dominant students in Lafayette Diocese Catholic Schools of Lousiana	Well matched on initial mental abiity and MRT pretests	**English Stanford**		
								Word Reading	+0.38	+0.26
								Paragraph meaning	+0.28	
								Vocabulary	+0.19	
								Word Study Skills	+0.23	
Bacon et al (1982)	Bilingual program	Matched control	4 & 5 yrs treatment (1st to 5th grade) and testing in 8th grade	53	1-5	Cherokee Indian students in Oklahoma	Well matched on control variables such as IQ and first language except for GPA & father's education, C>E	**English SRA Reading**		
								Cohort 1 (5 yrs) vs control	+0.73	+0.70
								Cohort 2 (4 yrs) vs control	+0.67	
Doebler & Mardis (1980)	Bilingual program	Matched control	1 yr	63	2	Choctaw students in MS	Well matched on their initial English proficiency	**English MAT**	+0.15	+0.15
Studies of Upper Elementary and Secondary Reading										
Covey (1973)	Bilingual program	Random assignment	1 yr	200	9	Spanish dominant students	Well matched on pretests	**English Stanford Diagnostic Reading**	+0.82	+0.82
Kaufman (1968)	Bilingual program	Random assignment	1& 2 yrs	139	7	Spanish dominant students in New York City	Initial CIA vocab and comprehension scores, language and non-language IQ, age, and Hoffman bilingual schedule scores were used as covariates	**2-yr school**		
								Word Meaning	E=C	
								Paragraph Meaning	E=C	
								1 yr school		
								Word Meaning	E>C	
								Paragraph Meaning	E>C	

Studies of Beginning Reading for Spanish-Dominant Students

Longitudinal Studies Using Random Assignment

Three small studies used random assignment to bilingual or immersion programs. Plante (1976) randomly assigned 55 Spanish-dominant, Puerto Rican children in a New Haven, Connecticut, elementary school to a "paired bilingual" model or to English-only instruction. Two cohorts of experimental students were taught all of their basic skills (reading, writing, math, science, social studies) in Spanish in first and second, or second and third grades. At the same time, they received English instruction designed to transition them to English-only instruction. After two years, second-graders in the program scored significantly higher on an English reading test than their English-only counterparts (ES=+0.62). Differences were not significant for third-graders, but still favored the bilingual group (ES=+0.24). Not surprisingly, the bilingual students also scored substantially higher on Spanish reading measures.

In a very similar study, Huzar (1973) randomly assigned two groups of Spanish-dominant, Puerto Rican children in Perth Amboy, New Jersey, to bilingual or English-only classes. One group (N=81) was in the study in first and second grades, and the other (N=79) was in the study in first through third grades. The experimental and control groups were well matched on IQ, SES, and initial achievement. As in the Plante (1976) study, students in the paired bilingual group had two teachers. One taught reading in Spanish for 45 minutes daily, while the other taught reading in English for the same amount of time. While the two groups did not differ after two years, children who were in the program for three years (grades 1-3) scored higher than the control group in English reading. Using the control group standard deviation the effect size would be +0.68, but experimental and control standard deviations are very different. Using a pooled standard deviation yields a more conservative ES=+0.31. This difference was not statistically significant (t=1.38).

J.A. Maldonado (1994) carried out a small, randomized study involving English language learners who were in special education classes in Houston. Twenty second- and third-graders with learning disabilities were randomly assigned to one of two groups. A bilingual group was taught mostly in Spanish for a year, with a 45-minute ESL period. During a second year, half of the instruction was in English, half in Spanish. In a third year, instruction was only in English. The control group was taught in English all three years.

Children were pretested on the CTBS and then posttested on the CTBS three years later. At pretest, the control group scored nonsignificantly higher than the bilingual group, but at posttest the bilingual group scored far higher. Using the means and standard deviations presented in the article, the effect size would be +8.33, but using the given values of t, the effect size is +2.21, a more credible result.

The Huzar (1973) and Plante (1976) studies are particularly important, despite taking place more than a quarter century ago. Both are multi-year experiments that, due to use of random assignment, can rule out selection bias as an alternative explanation for the findings. Both started with children in the early elementary grades and followed them for two to three years. Interestingly, both used

a model that would be unusual today, paired bilingual reading instruction by different teachers in Spanish and English, with transition to all-English instruction by second or third grade. The use of both Spanish and English reading instruction each day more resembles the experience of Spanish-dominant students in two-way bilingual programs (see Calderón & Minaya-Rowe, 2003) than it does typical transitional bilingual models, which delay English reading to second or third grade.

The J.A. Maldonado (1994) study of ELLs with learning disabilities found dramatically higher achievement gains for children transitioned over a three-year period from Spanish to English than for children taught only in English. Across these three randomized studies the median effect size is a substantial +0.43. These studies were, however, very small and two of them took place long ago, so they should not be considered by themselves as conclusive evidence in favor of bilingual education.

Longitudinal Studies Using Matching

One of the most widely cited studies of bilingual education is a longitudinal study by Ramirez et al. (1991) that compared Spanish-dominant students in English immersion schools to two forms of bilingual education: early exit (transition to English in grades 2-4) and late-exit (transition to English in grades 5-6). Schools in several districts were followed over four years. Immersion and early-exit students were well matched, but late-exit students were lower than their comparison groups in SES and their schools had much lower proportions of native English speakers. For these reasons, direct comparisons were not made between late-exit and other schools.

The comparison of early-exit, transitional bilingual education and English immersion is the important contribution of the Ramirez et al. (1991) study. It involved four schools, each of which provided both programs. The children in the two programs were well matched on kindergarten pretests, socioeconomic status, preschool experience, and other factors. They were tested on the English CTBS each spring in grades 1-3. In reading, the early-exit children scored significantly better than immersion students at the end of first grade. By third grade, these differences were in the same direction but were not statistically significant, controlling for premeasures.

The Ramirez et al. study was so important in its time that the National Research Council convened a panel in 1991 to review it and a study by Burkheimer et al. (1989). The panel's report (Meyer & Fienberg, 1992) supported the conclusions of the Ramirez et al. comparison of the early-exit and immersion programs in grades K-1.

Meyer and Fienberg (1992) did not support the conclusions of the Burkheimer et al. study on effects of various bilingual and immersion models, due to lack of clear comparisons of alternative treatments (among many other problems), and the Burkheimer et al. study was excluded from this review for similar reasons.

Saldate, Mishra, and Medina (1985) studied 62 children in an Arizona border town who attended immersion or bilingual schools. The children were individually matched on the Peabody Picture Vocabulary Test in first grade. At the end of second grade, the bilingual students scored nonsignificantly lower on the English Metropolitan Achievement Test (MAT) (ES=-0.29) and higher on the Spanish MAT (ES=+0.46). This was to be expected, as they had not yet transitioned to English in-

struction. At third grade, however, the bilingual students (who had now transitioned to English-only instruction) substantially outperformed the immersion students both in English (ES=+1.47) and in Spanish (ES=+6.40). This study's small size means that its results should be interpreted cautiously, especially as the number of pairs dropped from 31 to 19 between second and third grades.

Cohen (1975) compared two schools serving many Mexican Americans in Redwood City, California. One school was using what amounts to a two-way bilingual program, in that Spanish-dominant students and English-dominant students were taught in both Spanish and English. Three successive cohorts were compared at the two schools: grades 1-3, 1-2, and K-1. In each case, students were pretested and posttested on a broad range of English reading measures. In all cohorts, Mexican-American students were well matched on English and Spanish pretests. At posttest, there were no significant differences, adjusting for pretests. The data did not allow for computation of effect sizes. Similarly, a study in Corpus Christi, Texas, by J.R. Maldonado (1977) compared Mexican-American children in bilingual and English-only classes in grades 1-5, and found no differences at any grade, controlling for first-grade pretests. A study by Alvarez (1975) followed Mexican-American children in Austin, Texas, from first to second grades. There were no differences between children taught in English and those taught in English and Spanish.

In the mid-1970s, the American Institutes of Research (AIR) produced a series of reports on bilingual programs around the U.S. (Campeau, Roberts, Oscar, Bowers, Austin, & Roberts, 1975). These are of some interest, with one major caveat: The AIR researchers were looking for *exemplary* bilingual programs. They began with 96 candidates and ultimately winnowed this list down to eight. Programs were excluded if data were unavailable, not because they failed to show positive effects of bilingual programs. Nevertheless, these sites were chosen on their reputations for excellence, and a site would clearly be less likely to submit data if the data were not supportive of bilingual education. Also, the Campeau et al. (1975) evaluations were organized as successive one-year studies, meaning that pretests after the first treatment year (K or 1) are of little value. For reasons described earlier, one-year evaluations of bilingual education are likely to be biased against the bilingual group on early English measures. With these cautions in mind, the Campeau et al. (1975) studies are described below.

A study in Corpus Christi, Texas, evaluated a bilingual program in three schools. The kindergarten program made extensive use of both Spanish and English instruction and reading materials in both languages, but the emphasis was on Spanish (90% of the instruction). A control group, consisting of students in three different schools, was taught only in English. In the 1972-73 cohort, experimental and control classes were well matched on both English and Spanish measures.

At the end of kindergarten, the control group was slightly ahead on a standardized test of letters and sounds, but the bilingual group was slightly ahead on an English test of general ability. A second bilingual kindergarten cohort (1973-74) also slightly outscored the control group on general ability in English. The first-graders in 1973-74, who were the kindergarteners in the earlier analysis, ended the year with the bilingual students scoring 50% of a grade equivalent ahead of controls in SRA reading and substantially ahead of controls on general ability in English (ES=+0.45). They also were far ahead in Spanish ability. Because kindergarten pretests for these first graders were not

shown, however, these results should be interpreted with caution, as attrition over two years could have made the initially equivalent samples unequal.

Separate analyses within the bilingual groups found that the more years students in grades 1-3 had spent in bilingual education, the higher their scores. However, it is possible that the children who spent more years in bilingual education were simply less mobile than those who had fewer years in the program, and stable children typically have higher achievement than mobile ones.

A study in Houston followed three cohorts of students in seven bilingual and two immersion schools. On a kindergarten pretest of English ability the students in the immersion groups scored substantially higher in all three cases, but at the end of kindergarten the bilingual classes were substantially higher on the English ability test. Controlling for pretests, these differences were highly significant ($p<.001$). Bilingual first graders in the second cohort and first and second graders in the third cohort (the former kindergartners) consistently outscored students who were in the immersion program.

A study in Alice, Texas, compared Spanish-dominant bilingual and immersion students starting in kindergarten, for a two-year study. While kindergartners were comparable at pretest on English measures of general ability, bilingual students scored substantially higher on a Spanish ability test. At posttest (controlling for pretests), bilingual students scored substantially better in English reading at the end of first grade (after two years of bilingual education).

Two additional studies reported by Campeau et al. (1975) lasted only a year, and are discussed in the following section. Two more did not qualify for inclusion. One had no control group and the other lacked sufficient evidence of initial equality.

The Campeau et al. studies are far from representative of all bilingual programs, as they focused by design on exemplary programs. Several of these studies, however, did have well-matched control groups and met the review criteria. They must be taken seriously as additional evidence favoring bilingual programs that emphasize both Spanish and English materials and instruction in the early grades.

One-Year Studies

For the reasons discussed earlier, one-year studies of bilingual education, with posttests administered before children have transitioned from native-language instruction to English, are of limited value. For example, Carlisle and Beeman (2000) studied a school that had just begun a bilingual program in which 80% of instruction was in Spanish. They compared first graders in the first year of the new program to those the previous year, who had been taught entirely in English. Not surprisingly, the students in the bilingual program scored substantially better in Spanish, but there were few differences in English reading. Other one-year studies (e.g., Danoff et al. 1978) were excluded because their pretests were given after bilingual and immersion programs were underway, or because they compared nonequivalent groups (e.g., Rossell, 1990).

Two of the studies carried out by Campeau et al. (1975) had one-year durations. A one-year study in Kingville, Texas, did not present enough data for adequate evaluation in this review, but re-

ported significantly greater gains on English SRA achievement tests for bilingual kindergartners than for immersion kindergartners in all six classroom pairs assessed.

Another one-year study in Santa Fe, New Mexico, compared bilingual and immersion programs for Spanish-dominant students. Pre- and posttests are reported for each year but only first grade is interpretable, as pretests for other years had already been affected by the treatments. Parents chose to place their children in bilingual or English programs, and apparently parents of higher-achieving children chose the bilingual group, as pretest scores were higher in that group. However, the bilingual group also gained more in English reading than the English-only group. No standard deviations were given, so effect sizes for pretest differences and gains could not be computed.

Studies Involving Languages Other Than Spanish

Morgan (1971) carried out a study of almost 200 children from French-speaking parents in rural Louisiana. Existing groups of first graders, assigned to bilingual or monolingual classes, were followed for a year. In the bilingual classes, children were taught in both French and English. The two groups were virtually identical on English tests of mental abilities and readiness at the beginning of first grade. At the end, the children taught in the bilingual classes scored higher on four English reading measures, with a median difference of +0.26. Differences were significant on measures of word reading and paragraph reading, but not vocabulary or word study skills. It is important to note, however, that the children in this study were probably English proficient. Their parents may have spoken French at home, but both experimental and control students scored well at pretest on an English mental abilities test.

Bacon, Kidd, and Seaberg (1982) evaluated a bilingual program for Cherokee students in Northeastern Oklahoma. The program introduced Cherokee language and reading materials to supplement English materials. This was clearly a heritage language approach; children apparently spoke English, and 28% of them did not speak Cherokee. The experimenters tested children as eighth graders. Two groups of children had attended the bilingual school in grades 1-5 or 2-5. Matched children from other schools taught only in English were the control group. The groups were not well matched, however; the control group had many more girls, and higher IQ's, father's education, and grade point averages. On the eighth-grade tests all groups were nearly identical, but after using regression analyses to control for matching factors, the bilingual groups scored higher. However, one of the control variables was grade point average, which was higher in the control group, so the analysis may have overadjusted the control scores.

A one-year study of 63 Choctaw second graders in Mississippi compared a bilingual program in Choctaw to English-only instruction (Doebler & Mardis, 1980-81). There were no differences on an English reading measure, controlling for pretests.

Studies of Secondary Reading

Two qualifying studies evaluated programs that introduced Spanish-language instruction to ELLs in the secondary grades. Both of these used random assignment.

Covey (1973) randomly assigned 200 low-achieving Mexican-American ninth graders to bilingual or English-only classes. The experimental intervention is not described in any detail, but it clearly involved extensive use of Spanish in reading, English, and math. The groups' scores were nearly identical at pretest, but at posttest the bilingual students scored significantly better on the Stanford Diagnostic Reading Test (ES=+0.82).

Kaufman (1968) evaluated a program in which low-achieving Spanish-speaking seventh graders were randomly assigned to bilingual or English-only conditions in two New York junior high schools. One school participated in the program for a year and the other for two years. In the bilingual classes, students received three or four periods of Spanish reading instruction each week, while controls were in art, music, or health education. On standardized tests of word and paragraph meaning, there were no significant differences in the two-year school, but in the one-year school significant differences favored the bilingual group on one of two word meaning tests. Results of paragraph meaning tests favored the bilingual group, though not significantly. The data presented did not permit computation of effect sizes.

The secondary studies point to the possibility that providing native language instruction to low-achieving ELLs in secondary school may help them with English reading. This application is worthy of additional research.

Canadian Studies of French Immersion

There are several Canadian studies (e.g., Lambert & Tucker, 1972; Genesee & Lambert, 1983; Day & Shapson, 1988; Barik & Swain, 1978) that have played an important role in debates about bilingual education. These are studies of French immersion programs, in which English speaking children are taught entirely or primarily in French in the early elementary years. Rossell and Baker (1996) emphasize these studies as examples of "structured English immersion," the approach favored in their review. However, Willig (1985) and other reviewers have excluded them. These studies do not meet the inclusion standards of this review because the anglophone children are learning a useful second language, not the language for which they will be held accountable in their later schooling. Although most of the studies took place in Montreal, the children lived in English-speaking neighborhoods, and attended schools in an English system. The purpose of bilingual education is to help children succeed in the language in which they will be taught in the later grades, but the French immersion children in Canada are headed to English secondary schools. Further, these studies all involve voluntary programs, in which parents wanted their children to learn French, and the children in these studies were generally upper middle class, not disadvantaged.

Because French immersion programs were voluntary, children who did not thrive in them could be and were routinely returned to English-only instruction. This means that the children who complete French immersion programs in Canada are self-selected, relatively high achievers. Most importantly, the "bilingual" programs to which French immersion is compared are nothing like bilingual education in the U.S. At most, children receive 30 to 40 minutes daily of French as a second language, with far less time in French reading instruction than a U.S. student in a bilingual program would receive in English during and after transition (see Genesee & Lambert, 1983). Yet in many

studies, English comparison groups were not learning French at all. In the widely cited study by Lambert and Tucker (1972), anglophones in French immersion classes were compared to anglophones taught only in English, and to francophones taught only in French. Ironically, studies of this kind, cited by Rossell and Baker (1996) as comparisons of immersion and bilingual education, are in fact comparisons of immersion and *monolingual* education. If they existed, Canadian studies of, say, Spanish speakers learning French in francophone schools in Quebec or English in anglophone schools in the rest of Canada would be relevant to this review, but studies of voluntary immersion programs as a means to learn French as a second language are only tangentially relevant.

While the Canadian immersion studies are not directly relevant to the question of the effectiveness of bilingual programs for ELLs learning the societal language, they are nevertheless interesting in gaining a broader understanding of the role of native language in foreign language instruction. As a group, these studies are of high methodological quality. Quite in contrast to U.S. studies, however, the focus of the Canadian studies is whether or not French immersion harms the English language development of native English speakers. It is taken as obvious that French all day will produce more facility in French than 30 to 40 minutes daily in second language classes.

Lambert and Tucker (1972) carried out the foundational study of French immersion in Canada. It compared anglophone children taught completely in French from kindergarten and first grade, with some English instruction in grades 2-4, to matched anglophone children taught in English and to francophone children taught in French. At the end of first grade, immersion children scored far below children taught in English on English reading measures. And, while their spoken French was much worse than the French controls, their French reading was as good as that of the native speakers. A study of a second cohort found similar results at first grade. At second grade, however, the immersion students had almost caught up to the English-only students, and there were no differences in third or fourth grade in either English (compared to English-only anglophones) or French (compared to French-only francophones). A followup to grades 5-6 found the same patterns (Bruck, Lambert, & Tucker, 1977)

The finding of no differences was taken as a vindication of French immersion, as the anglophone children suffered no loss in English reading and gained fluent reading and speaking skills in an important second language. Because the comparison students were taught in only one language, however, there is no "bilingual" group to which immersion could be compared.

Other French immersion studies followed a similar paradigm. Barik and Swain (1975) studied a program in Ottawa, which also found similar second grade English reading performance for anglophone children taught entirely in French in grades K-1, with 60 minutes daily of English instruction in second grade, compared to anglophone children taught only in English. There were no differences in English reading by the end of second grade. Another Ontario study by Barik, Swain, and Nwanunobi (1977) compared a "partial French immersion" program (essentially, a paired bilingual program with 50% of instruction in each language) to English-only instruction. The English-only students performed better in English reading through third grade, but in grades 4 and 5 the two groups were similar, and the partial immersion students were fluent in French.

A study by Genesee, Lambert, Sheiner, and Tucker (1983) evaluated a trilingual immersion approach, in which anglophone children in Montreal were immersed in Hebrew as well as French.

Compared to a Hebrew-emphasis school that had one-hour French as a second language classes and three hours of Hebrew per day, the immersion students performed at similar levels in English reading by the end of second grade. Not surprisingly, the immersion school scored better in French than the Hebrew-emphasis school, but scores in Hebrew were mostly comparable. A followup through fifth grade found similar outcomes (Genesee, Lambert, & Tucker, 1977), as did another study of "trilingual education" (Genesee & Lambert, 1983).

Two studies compared early French immersion (starting in grades K-1) to delayed immersion (starting in grade 4). Genesee, Holobow, Lambert, and Chartrand (1989) found that in fifth grade, all groups scored equally in English reading, but the early immersion and French-only anglophones scored better than the delayed immersion students in French reading. A study in British Columbia also compared early and delayed immersion and found that by seventh grade, both had similar effects on French and English reading (Day & Shapson, 1988).

Overall, the Canadian studies paint a consistent picture. At least for the overwhelmingly middle-class students involved, French immersion had no negative effect on English reading achievement, and it gave students facility in a second language. The relevance to the U.S. situation is in suggesting that similar second-language immersion programs, as well as two-way bilingual programs for English proficient children, are not likely to harm English reading development. However, the relevance of these studies to any context in which the children of immigrants are expected to learn the language that will constitute success in their school and in the larger society is unclear.

Comparisons of Paired Bilingual and Transitional Bilingual Programs

As noted earlier, many of the programs with the strongest positive effects for English language learners used a paired bilingual approach, in which children were taught reading in both English and their native language at different times each day from the beginning of their schooling. This approach contrasts with transitional bilingual education (TBE) models in which children are first taught to read primarily in their native language, and only then transitioned gradually to English-only instruction. Two studies have compared reading outcomes of these two bilingual approaches.

A longitudinal study by Gersten and Woodward (1995) initially favored paired bilingual instruction over TBE, but later found them to be equivalent. This study was carried out with Spanish-dominant ELLs in 10 El Paso elementary schools. Five schools used a program in which all subjects were taught in English, but Spanish instruction was also provided, for 90 minutes daily in first grade declining to 30 minutes a day in fourth grade. The transitional bilingual program involved mostly Spanish instruction with one hour per day for ESL instruction, with gradual transition to English completed in the fourth or fifth grade. The children were well matched demographically on entry to first grade, and scored near zero on a measure of English language proficiency. In grades 4, 5, 6, and 7, Iowa Tests of Basic Skills were compared for the two groups. On Total Reading, the paired bilingual students scored significantly higher than the transitional bilingual students in fourth grade (ES=+0.31), but the effects diminished in fifth grade (ES=+0.18), and were very small in sixth (ES=+0.06) and seventh grades (ES=+0.08). Tests of language and vocabulary showed similar patterns. This pattern is probably due to the fact that the transitional bilingual students had not completed their transition to English in fourth and fifth grades. When they had done so, by sixth grade, their reading performance was nearly identical.

A one-year study of Spanish-dominant kindergartners by Pena-Hughes and Solis (1980) could not be located, but information provided by Willig (1985) indicated that it found a strong advantage of a paired bilingual approach over a TBE model. As the paired bilingual program provided more time in English instruction, however, a longer study would be needed to establish the relative effects.

Research comparing alternative bilingual models is inconclusive, but nothing suggests that it is harmful to children's reading performance to introduce both native language and English reading instruction at different times each day.

Conclusions: Language of Instruction

The most important conclusion from research on language of instruction is that there are far too few high-quality studies of this question. Willig (1985) and Rossell and Baker (1996) agree on very little, but both of these reviews call for randomized, longitudinal evaluations to produce a satisfying answer to this critical question. Of course, many would argue that randomized evaluations are needed on most important questions of educational practice (see, for example, Mosteller & Boruch, 2002; Slavin, 2003), but in bilingual education, this is especially crucial due to the many inherent problems of selection bias in this field. Further, this is an area in which longitudinal, multi-year studies are virtually mandatory, to track children initially taught in their native language through their transition to English. Finally, while randomized, longitudinal studies of this topic are sorely needed, there are simply too few experimental studies of all kinds, including ones with matched experimental and control groups.

With these concerns in mind, however, research on language of instruction does yield some important lessons at least worthy of further study. First, there are three randomized, multi-year evaluations of bilingual programs for beginning reading (Huzar, 1973; Plante, 1976; Maldonado, 1994). All three found effect sizes that moderately to strongly favored bilingual over immersion models. Yet all three used strategies that are quite different from bilingual models common in recent years. Both Huzar (1973) and Plante (1976) used paired bilingual models in which children were taught reading in both English and Spanish, at different times every day. In the study by J.A. Maldonado (1994), the bilingual classes were taught to read in Spanish for one year, in Spanish and English in the second year, and in English in the third year, a more rapid transition than in typical transitional bilingual programs. These studies hold out an intriguing possibility that English language learners may learn to read best if taught *both in their native language and in English*, from the beginning of formal instruction. Rather than confusing children, as some have feared, reading instruction in a familiar language may serve as a bridge to success in English, as decoding, sound blending, and generic comprehension strategies clearly transfer among languages that use phonetic orthographies, such as Spanish, French, and English (see August, 2002; August & Hakuta, 1997; Fitzgerald, 1995; Garcia, 2000). Two studies comparing paired bilingual to transitional bilingual programs were far from conclusive, but did not provide evidence indicating an advantage of transitional approaches.

Looking past the small number of randomized experiments, outcomes of multi-year studies with pretests available before treatments began show mixed results, but most such studies favor bi-

lingual over immersion approaches. Among studies that met the criteria for inclusion, none significantly favored immersion programs, but some found no differences.

Only two studies of secondary programs met the inclusion criteria, but both of these were very high quality randomized experiments. Covey (1973) found substantial positive effects of Spanish instruction for low-achieving ninth graders, while Kaufman (1968) found mixed, but slightly positive, effects of a similar approach with low-achieving seventh graders.

As noted previously, research on language of instruction may suffer from publication bias, the tendency for journals to publish only articles that find significant differences. However, dissertations and technical reports less likely to suffer from publication bias also tended to favor bilingual programs.

The findings of this review are at significant variance from those of both Rossell and Baker (1996) and Willig (1985), the most widely cited quantitative syntheses in this area. They correspond closely, however, to the findings of a meta-analysis by Greene (1997), who also concluded that most methodologically adequate studies, including all of those using random assignment, favored bilingual approaches.

EFFECTIVE READING PROGRAMS FOR ENGLISH LANGUAGE LEARNERS

Many reviewers of research on programs for English language learners (e.g., August & Hakuta, 1997; Brisk, 1998; Christian & Genesee, 2001; Goldenberg, 1996; Secada et al., 1998) have concluded that researchers need to focus more on the *quality* of instruction for English language learners, rather than continuing to argue primarily about language of instruction. There are surely better and worse bilingual programs as well as better and worse immersion programs. In fact, there is a significant body of research that offers insights into effective reading programs for English language learners. This research is reviewed in the following sections.

How Do English Language Learners Acquire Reading Skills?

Researchers who study reading instruction for English language learners have asked whether these children learn to read in the same way as those who are proficient in English, or whether there are different dynamics involved. In general, reviewers of this literature have concluded that the factors that lead to reading in English language learners are similar to those for their English-proficient classmates. For example, oral English proficiency is highly correlated with (or predictive of) English reading (August, 2002). English vocabulary, however, is also highly predictive of reading performance among English-proficient children. The National Reading Panel (2000) systematically reviewed research on early reading, and identified five major elements that contribute to early reading success among English-proficient children: phonemic awareness, phonics, vocabulary, comprehension, and fluency. Yet these same factors have also been linked to English reading success for English language learners (see August, 2002; Baker & Gersten, 1997; Garcia, 2000; Fitzgerald, 1995; Gersten &

Geva, 2003). This does not mean that no accommodations are necessary for English language learners, but it does suggest that with allowances for the language issues themselves, effective reading instruction for English language learners may be similar to effective instruction for English-proficient children, whether the ELLs are first taught in their native language or in English.

Review Methods and Criteria for Inclusion

Review methods for studies of reading programs for English language learners were the same as for language of instruction, with a few differences in criteria for inclusion. These are as follows:

1. The studies compared children taught reading in classes using a given reading method and those in control classes using standard methods. In contrast to the previous section, language of instruction is the same in experimental and control groups.
2. Random assignment or matching with appropriate adjustments for any pretest differences had to be used, as in the previous section.
3. Subjects were English language learners in elementary or secondary schools in English-speaking countries, as described previously.
4. The dependent measures included quantitative measures of reading performance, as described previously. In this section, however, measures of performance in languages other than English are reported if the study compared alternative strategies for teaching reading in that language.
5. A minimum treatment of 12 weeks was required. This is shorter than the one-year duration specified for studies of language of instruction, as the problem of having to wait until transition has been completed does not exist in this set.

Studies of Beginning Reading Programs

It is in the earliest years of formal education that children define themselves as learners, largely on the basis of reading success. The early elementary years are of particular importance for English language learners, as this is the time when they are most likely to be struggling both to learn a new language and to learn to read. Perhaps because of this, the largest number of methodologically adequate studies have focused on the early elementary grades. Studies in this section are ones in which the treatments begin in kindergarten, first, or second grades.

There were 11 studies of beginning reading that fully met the criteria outlined above. Most studies of reading approaches for English language learners lack control groups or objective measures, do not document or control for pretest differences, or are very brief. The main characteristics and findings of the qualifying studies are summarized in Table 2.

TABLE 2
Beginning Reading Programs: Descriptive Information and Effect Sizes for Qualifying Studies

Study	Intervention Description	Design	Duration	N	Grade	Sample Characteristics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Success For All										
Nunnery et al (1997)	SFA-Bilingual	Matched control	1 yr	298	1	Spanish-dominant students across 30 schools with bilingual programs in Houston TX	Fairly well matched on demographic and well matched on pretest. C>E; ES=-0.08	**Spanish Woodcock**		
								Medium Implementation		
								Word Identification	+0.20	+0.22
								Word Attack	+0.30	
								Passage Comprehension	+0.22	
								Low Implementation		
								Word Identification	+0.27	+0.22
								Word Attack	+0.22	
								Passage Comprehension	+0.17	
Livingston & Flaherty (1997)	SFA-Bilingual	Matched control	3 yrs	6 schools (3 E & 3 C)	1-3	Spanish-dominant bilingual students in CA	Well matched on demographics and PPVT pretests, median ES across cohorts=+0.21	**Spanish Woodcock**		
								92 cohort		+0.97 (Grade 1) +0.44 (Grade 2) +0.03 (Grade 3)
								Grade 1	+0.97	
								Grade 2	+0.45	
								Grade 3	+0.03	
								93 cohort		
								Grade 1	+0.72	
								Grade 2	+0.43	
								94 cohort		
								Grade 1	+1.41	
	SFA-English Language Development Adapation					Spanish-dominant ESL students in CA	Well matched on demographics and PPVT pretests, median ES across cohorts=+0.26	**English Woodcock**		
								92 cohort		+1.36 (Grade 1) +0.46 (Grade 2) -0.09 (Grade 3)
								Grade 1	+1.36	
								Grade 2	+0.19	
								Grade 3	-0.09	
								93 cohort		
								Grade 1	+1.32	
								Grade 2	+0.72	
								94 cohort		
								Grade 1	+1.40	
	SFA-English Language Development Adapation					Other ESL students in CA	Well matched on demographics and PPVT pretests, median ES across cohorts= -0.10	**English Woodcock**		
								92 cohort		+0.24 (Grade 1) +0.37 (grade 2) +0.05 (Grade 3)
								Grade 1	+0.24	
								Grade 2	+0.25	
								Grade 3	+0.05	
								93 cohort		
								Grade 1	+0.96	
								Grade 2	+0.49	
								94 cohort		
								Grade 1	+0.05	

Study	Intervention Description	Design	Duration	N	Grade	Sample Characteristics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Slavin & Madden (1995)	SFA-English Language Development Adapation	Matched control	5 yrs	50	K	Asian students in 2 schools in Philadelphia	Well matched on overall achievement level, poverty, and other variables	**English Woodcock**	**Grade 4**	
								Word Identification	+1.54	+1.49
Slavin & Yampolsky (1991)								Word Attack	+1.49	
								Passage Comprehension	+0.62	
								English Woodcock	**Grade 5**	
Slavin, Leighton & Yampolsky (1990)								Word Identification	+1.40	+1.33
								Word Attack	+1.33	
								Passage Comprehension	+0.75	
Ross et al (1998)	SFA-English Language Development Adapation	Matched control	1 yr	540	1	Stratum 1 (Low SES): 50% Hispanic and 81% free lunch; Stratum 2 (Mid SES): 27% Hispanic and 53% free lunch	Well matched on demographics and fairly well on pretests: Stratum 1 (Low SES) E>C; ES=+0.54; Stratum 2 (Mid SES), E>C, ES=+0.22	**English Woodcock**	**Low SES**	
								Word Identification	+0.39	+0.39
								Word Attack	+0.59	
								Passage Comprehension	+0.39	
								Durell	+0.31	
								English Woodcock	**Mid SES**	
								Word Identification	+0.63	+0.53
								Word Attack	+1.07	
								Passage Comprehension	+0.43	
								Durell	+0.32	
Hurley et al (2001)	SFA	Compared gains to the state mean for Hispanic students	4 yrs	95 SFA schools	(K-2)--> (3-5)	Hispanic students in TX	Well matched on initial TAAS reading scores	**English TAAS Reading (Grade 3-5)**	+0.28*	+0.28* (ES from school means, not individual scores)
Other Programs										
Becker & Gersten (1982)	Direct instruction	Matched control	follow up study--2 yrs after the treatment	225	K-3	Hispanic ELL students in Uvalde, TX	Well matched on demographics	**English WRAT Reading**	**Across 2 grades**	
								Level II	+0.44	+0.21
								Level I	+0.50	
								Mean	**+0.47**	
								English MAT		
								Word Knowledge	+0.11	
								Reading	+0.21	
								Total Reading	+0.16	
								Mean	**+0.16**	
Gersten (1985)	Direct instruction	Matched control	8 mos	~35	1-2	Asian ELL students	Similar on LAS scores for cohort 1 (C>E) and cohort 2 (C>E) and fairly well matched on demographic	**English CTBS Reading**		
								Experimental	**75%**	E>C
								Control	**19%**	
								English CTBS Language		
								Experimental	**71%**	E>C
								Control	**44%**	

Study	Intervention Description	Design	Duration	N	Grade	Sample Characteristics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Stuart (1995)	Phonetic program (Jolly Phonics) vs Literature-based program (Big Books)	Matched control	12 wks	112	K	Sylheti-dominant students in London	Well matched on demographics but not on pretests; JP>BB; ES=+0.88 on phonics knowledge pretests; JP>BB; ES=+0.70 on reading and writing pretests	**English**		
								Phoneme awareness (5 measures)	+0.70	Immediate tests: +0.88
								Delayed tests (1 year later)	+0.16	
								Reading and Spelling (5 measures)	+1.06	Delayed tests: +0.34
								Delayed tests (1 year later)	+0.52	
Escamilla (1994)	Reading Recovery in Spanish	Matched control	7 mos	46	1	Spanish-dominant bilingual students in Arizona	Well matched on Spanish Aprenda, but on Spanish observation survey, C>E, median ES=-0.43 across four measures	**Spanish**		
								Spanish Aprenda	+0.30	+0.30
								Spanish Observation Survey (6 measures)	+0.84	+0.84
Gunn et al., (2000)	Small group tutoring using Direct Instruction	Random assignment	2 yrs	122	K-4	Low-achieving Spanish-dominant students in rural Oregon	Well-matched on English Woodcock-Johnson, oral reading fluency	**English Woodcock**		
								Year 1 -Letter Word +0.22 -Word Attack +0.70 -Oral Reading Fluency +0.16		Year 1 +0.22
								Year 2 -Letter Word +0.46 -Word Attack +0.91 -Oral Reading +0.43 -Vocabulary +0.44 -Comprehension +0.43		Year 2 +0.44
Goldenberg et al (1990)	Use of teacher-created booklets at home and at school	Matched control	8 mos	48	K	Spanish-dominant students in Southern CA	Similar on Bilingual Syntax Measure and free lunch	**Spanish**		
								13 measures of early literacy development	+0.83	+0.83

Success for All

Among the studies that did fully meet the inclusion criteria, five evaluated the Success for All program (Slavin & Madden, 1999, 2001). Success for All is a comprehensive reform model that focuses school resources and energies on seeing that all children succeed in reading from the beginning of their time in school. It provides schools with well-structured curriculum materials emphasizing systematic phonics in grades K-1 and cooperative learning, direct instruction in comprehension skills, and other elements in grades 2-6. It provides extensive professional development and followup for teachers, frequent assessment and regrouping, one-to-one tutoring for children who are struggling in reading, and family support programs. A full-time facilitator helps all teachers implement the model.

For English language learners, Success for All has two variations. One is a Spanish bilingual program, *Exito para Todos*, which teaches reading in Spanish in grades 1-2 and then transitions them to English-only instruction, usually starting in third grade. The other is an English language development (ELD) adaptation, which teaches children in English with appropriate supports, such as vocabulary development strategies linked to the words introduced in children's reading texts. In both adaptations, children at the lowest levels of English proficiency usually receive separate instruction during a time other than the reading period to help develop their oral language skills.

Studies of Success for All with English language learners have generally compared children taught using the Spanish adaptation to other children taught in Spanish, or have compared the ELD adaptation to other ELD English reading programs.

Success for All: Spanish Bilingual Adaptation (*Exito para Todos*)

California (Bilingual). Researchers at the Southwest Educational Research Laboratory (now part of WestEd) conducted a three-year longitudinal study involving three California elementary schools and three matched controls. They pooled data across the schools in four categories: English-dominant students, Spanish-dominant students taught in Spanish (*Éxito Para Todos)*, Spanish-dominant students taught in English, and speakers of languages other than English or Spanish taught in English. Three cohorts were followed. Data for a 1992 cohort were reported for grades 1, 2, and 3; for 1993, grades 1 and 2; and for 1994, grade 1 only.

Students in the two *Éxito Para Todos* schools in California scored higher on the Spanish Woodcock than controls at every grade level in all three cohorts (Livingston & Flaherty, 1997, Dianda & Flaherty, 1995). Median effect sizes across cohorts averaged +0.97 for first graders, +0.44 for second graders, and +0.03 for third graders. The analyses for second and third graders understate the magnitude of the differences. The authors note that in line with district and program policies, students initially taught in Spanish were transitioned into English instruction as soon as they demonstrated an ability to succeed in English. Because of their success in Spanish reading, many more *Éxito Para Todos* than control students were transitioned during second and third grades. Therefore, the highest-achieving experimental students were being removed from the Spanish sample, reducing the mean for this group.

Houston (Bilingual). The largest study of *Éxito Para Todos* took place in the Houston Independent School District (HISD). Both Spanish and English forms of Success for All were studied (see Nunnery, Slavin, Madden, Ross, Smith, Hunter, & Stubbs, 1997).

The Houston study was unusual in several ways. In contrast to other studies (and to standard practice in implementing Success for All in dissemination sites), schools were allowed to choose how completely to implement the program. They could choose to implement all program elements, the reading program and tutoring without other elements, or just the reading program. The intention was to compare outcomes according to degree of implementation.

The bilingual portion of the study compared first graders in 20 schools implementing *Éxito Para Todos* to those in 10 matched schools also using Spanish bilingual instruction. Children were assessed on three scales from the Spanish Woodcock: Word Identification, Word Attack, and Passage Comprehension. Ten children were selected at random from each school; after missing data were removed, there were 298 Spanish-dominant students across the 30 schools with bilingual programs.

The Success for All schools were grouped into three categories of implementation—high, medium, or low—based on such implementation categories as whether the school had a full-time, part-time, or no facilitator, the number and certification status of tutors, and the existence of a family support team. Among the bilingual schools, no school fell into the "high" category, primarily because few had certified teachers working as bilingual tutors. The medium-implementation schools, however, had many more paraprofessional tutors and were much more likely to have a full-time facilitator and a family support team than were the low-implementation schools. Otherwise, both sets of schools were very similar to each other and to bilingual programs in comparison schools. The Spanish-dominant SFA students were somewhat more impoverished than those in comparison schools, and had somewhat higher mobility.

School-level comparisons showed significant differences ($p<.05$) between both categories of SFA schools and comparison schools on Word Identification and Word Attack, and a marginally significant difference ($p<.06$) between medium implementation schools and controls on Passage Comprehension. Overall, median student-level effect sizes in comparison to controls were +0.22 for both medium implementers and for low implementers.

The study of the English version of Success for All also found positive achievement effects for the high-implementing and medium-implementing schools, but not for low implementers. These schools served primarily African-American and Hispanic children (Nunnery et al., 1997).

Philadelphia (Bilingual). The bilingual version of Success for All, *Éxito Para Todos*, was first implemented at Fairhill Elementary School, an inner-city Philadelphia school, starting in 1992 (see Slavin & Madden, 1994). Fairhill served 694 students of whom 78% were Hispanic (primarily from Puerto Rico) and 22% were African-American. A matched comparison school was also selected. The two schools were very similar in total enrollment, percent of Hispanic and African-American students, and historical achievement levels. The schools were also similar in the percent of students receiving bilingual instruction. In both schools about half of all students were in the bilingual program in first grade. Nearly all students in both schools qualified for free lunches.

A misunderstanding about the instruction provided by the control school changed the meaning of this experiment from its original intention. The control group's reading program was described by the district as a bilingual model emphasizing native language instruction. However, it turned out that the control group's "bilingual" approach was more of a sheltered English model, with very little instruction in Spanish. This made the Fairhill experiment a comparison of *Éxito Para Todos* (in Spanish) to a sheltered English control group, mixing language of instruction with method of instruction. For this reason, the study is not shown in Table 2. Its findings are interesting, however.

All students defined by district criteria as Limited English Proficient (LEP) at Fairhill and its control school were pretested at the beginning of first grade on the Spanish Peabody Picture Vocabulary Test (PPVT). Each following May, these students were tested by native Spanish speakers on three scales of the Spanish Woodcock: Letter/Word Identification, Word Attack, and Passage Comprehension. Starting in third grade, almost all children had transitioned to English instruction, so students were assessed on the corresponding English Woodcock scales as well.

A check for pretest differences on the Spanish PPVT found that there were differences in favor of the experimental group ($p < .03$). PPVT scores were therefore used as covariates in all analyses of covariance (ANCOVA). Not surprisingly, Fairhill students performed far better than control students on all three Spanish measures ($p<.001$; median ES=+2.53). More significant, however, were the differences in English reading performance. Fairhill students scored higher than control students on all three English reading measures in the third grade, although the differences were only statistically significant on Word Attack ($p<.05$; ES=+0.65). This finding contributes evidence that well-structured instruction in Spanish followed by systematic transition to English can lead to enhanced English reading, but the small study size and confounding of teaching methods with language of instruction make this study speculative rather than conclusive.

Success for All: English Language Development Adaptation

Philadelphia (ELD). The first evaluation of the English language development (ELD) adaptation of Success for All began at Philadelphia's Francis Scott Key Elementary in 1988 (see Slavin & Yampolsky, 1991; Slavin, Leighton, & Yampolsky, 1990; Slavin & Madden, 1995). Sixty-two percent of Key's students were from Asian backgrounds, primarily Cambodian. Nearly all of them entered the school in kindergarten with little or no English. The remainder of the school was divided between African American and White students. The school is in an extremely impoverished neighborhood in South Philadelphia.

The program at Francis Scott Key was evaluated in comparison to a matched Philadelphia elementary school. The two schools were very similar in overall achievement level and other variables. Thirty-three percent of the comparison school's students were Asian (mostly Cambodian), the highest proportion in the city after Key. The percentage of students receiving free lunch was very high in both schools, though higher at Key (96%) than at the comparison school (84%).

The data reported are for all students in grades 4-5 in Spring, 1995 (Slavin & Madden, 1995). With the exception of transfers, all students had been in the program since kindergarten. All students in grades 4-5 were individually administered three scales from the Woodcock Language Proficiency

Battery (Woodcock, 1984): Word Identification, Word Attack, and Passage Comprehension. Asian Success for All students at both grade levels performed substantially better than Asian control students. Differences between Success for All and control students were statistically significant on every measure at every grade level ($p<.001$). Median grade equivalents and effect sizes were computed across the three Woodcock scales. On average, Asian Success for All students exceeded Asian control students in reading grade equivalents by 2.9 years in fourth grade (Median ES = +1.49), and 2.8 years in fifth grade (Median ES = +1.33). Asian Success for All students were reading about a full year above grade level in fourth (GE = 5.8) and fifth grades (GE = 6.8), while similar control students averaged 1.9 years below grade reading level in fourth grade and 1.8 years below grade level in fifth grade.

California (ELD). The three-year California study (Livingston & Flaherty, 1997; Dianda & Flaherty, 1995) included data on English language learners taught in English. These included both students in one Modesto school that did not have a bilingual program, as well as ELLs in the two schools (one in Modesto and one in Riverside) who were speakers of languages other than English or Spanish.

Results for Spanish-dominant students taught in English show strong impacts for first graders (ES = +1.36), smaller ones for second graders (ES = +0.46), and no differences for third graders (ES = -0.09). Again, the transitioning of successful students out of ESL classes reduced the apparent differences by third grade (because the highest achieving students were no longer receiving ESL services).

Results for speakers of languages other than English or Spanish (taught in English) were similar to those for Spanish-dominant ESL students, except that there were no differences for the 1994 first grade cohort. Averaging across cohorts, effect sizes were +0.24 for first graders, +0.37 for second graders, and +0.05 for third graders (Livingston & Flaherty, 1997; Dianda & Flaherty, 1995).

Arizona (ELD). Another study of the ELD adaptation of Success for All in schools serving many students acquiring English was conducted in an Arizona school district (Ross, Smith, & Nunnery, 1998). This one-year study compared first graders in two Success for All schools, in three schools using locally developed Title I schoolwide projects, and in one school using Reading Recovery. Two strata of schools were compared. Stratum 1 consisted of very impoverished schools, in which 81% of students received free lunch and 50% were Hispanic. Stratum 2 consisted of less impoverished schools, in which 53% of students received free lunch and 27% were Hispanic.

Students were pretested on the English Peabody Picture Vocabulary Test (PPVT) and then posttested on the Woodcock Word Identification, Word Attack, and Passage Comprehension scales, and the Durrell Oral Reading Test. Analyses of covariance compared schools in each stratum to the other two schools in the same stratum, controlling for PPVT pretests.

In the highest-poverty schools (Stratum 1), Hispanic Success for All students scored significantly higher than the average of students in the two locally developed schoolwide projects on all measures (median ES = +0.39). Hispanic first graders in Success for All schools averaged at grade level (mean=1.80), but the comparison groups were below grade level on all measures (mean =1.45). Results were similar for the less impoverished schools (Stratum 2); Success for All Spanish-

dominant students scored significantly higher than those in the locally developed schoolwide project and the Reading Recovery school taken together (median ES = +0.53). The Reading Recovery and local schoolwide project schools did not differ significantly on any measure. Among children who received tutoring, Success for All students scored substantially higher than those tutored using Reading Recovery (median ES=+1.04).

Separate analyses for limited English proficient students in both strata found results similar to those for the full samples.

Texas Statewide Evaluations of Success for All. Hurley, Chamberlain, Slavin, and Madden (2001) reported an analysis of data from the Texas Assessment of Academic Skills (TAAS), comparing reading gains (from the year schools began to implement Success for All [1994 to 1997] to 1998) by all 111 Success for All schools in the state to those made by students throughout Texas. The comparisons involving Hispanic students are relevant to this review. Note that while the TAAS data were for grades 3-5, most of the students had been in the program three to four years, meaning that they had begun in grades K-2.

Ninety-five of the Success for All schools had enough Hispanic students in grades 3-5 to be included in the analysis. Analyzing at the school level, their TAAS reading gains were significantly greater ($p<.007$) than those for Hispanic students in the state as a whole. Hispanic students in the SFA schools and state means for Hispanic students were similar in the year before SFA was introduced. The effect size for school means was +0.28 (note: effect sizes from school means should not be compared with those from individual scores). An update of this analysis found that from 1999 to 2002, Hispanic students in Texas Success for All schools gained 10.8 percentage points in students passing TAAS, while Hispanic students in the state as a whole gained 5.0 percentage points. Also, among 48 schools reporting results for limited English proficient (LEP) students (of all language backgrounds), Success for All students gained 16.9 percentage points on the English TAAS from 1999 to 2002, while LEP students in the state as a whole gained 5.7 percentage points. The Texas comparisons combine students who were initially taught in English and those who were initially taught in Spanish.

Success for All: Conclusions

The effects of Success for All on the achievement of English language learners are not entirely consistent, but in general they are substantially positive. In all schools implementing *Éxito Para Todos*, the Spanish bilingual adaptation of Success for All, effect sizes for first graders on Spanish assessments were very positive. For students in the ELD adaptation of Success for All, effect sizes for all comparisons were also positive, for Cambodian students in Philadelphia (Slavin & Madden, 1995) as well as Mexican-American students in California and Arizona (Livingston & Flaherty, 1997; Ross et al., 1998). On longer-term measures, results were more mixed. There were no differences between experimental and control third graders in the Livingston and Flaherty (1997) study, but this is apparently due to the disproportionate transfer of high-achieving third graders out of the ESL classes. In contrast, long-term effects in a four-year study of the ELD adaptation (Slavin & Madden, 1995), and a four-year Texas statewide study of 95 schools serving Hispanic students (Hurley et al., 2001), found positive program impacts over the long term.

Direct Instruction

Direct Instruction (DI), or Distar (Adams & Engelmann, 1996), is a reading program that starts in kindergarten with very specific instructions to teachers on how to teach beginning reading skills. It uses reading materials with a phonetically controlled vocabulary, rapidly paced instruction, regular assessment, and systematic approaches to language development. DI was not specifically written for English language learners or Latino students, but it is often used with them.

The most important evaluation of DI was the Follow Through study of the 1970s, in which nine early literacy programs were evaluated (Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977). In sites throughout the U.S., matched experimental and control schools were compared on various measures of reading.

One of the sites was in Uvalde, Texas, which primarily served Hispanic students. Becker and Gersten (1982) carried out a followup of the Follow Through study when the children who had experienced the treatments in grades K-3 were in grades 5-6. They found that the Uvalde DI students, who were well matched on demographic factors with their control group, scored substantially better than the controls. Effect sizes averaged +0.47 for two scales of the individually administered WRAT and +0.16 across three Metropolitan Achievement Test (MAT) subscales, for a median across five tests in two grades of ES=+0.21.

Gersten (1985) evaluated DI as part of a structured immersion program for limited English proficient students who spoke a variety of Asian languages. In addition to the DI beginning reading program, the structured immersion model emphasized English at a level understood by the students, occasional translation, preteaching of vocabulary, and direct teaching of the structure of the English language. Students in a matched control group participated in programs whose characteristics were not described, but which also primarily taught in English.

The testing strategy was far from satisfactory, but was apparently biased against the DI children. Following district guidelines, teachers in the control schools could choose to excuse from testing students who they felt to be performing below grade level. It is not stated how many students were excused for this reason. None of the DI students was excluded.

Across two cohorts, 75% of DI students scored at, or above, grade level on the CTBS Total Reading Scale at the end of two years, while only 19% of comparison students were at or above grade level ($p<.001$).

A third study of DI, by Gunn, Biglan, Smolkowski, and Ary (2000), used a small-group tutorial model, and is therefore discussed later in this section.

Jolly Phonics (Systematic Phonics Instruction)

Stuart (1999) carried out an evaluation of Jolly Phonics, an English phonetic kindergarten reading program, in five London primary schools. This program was compared to a big books program emphasizing teaching by drawing children's attention to letters and words in popular children's stories. The subjects were mostly English language learners, and among these most were speakers of Sylheti (a language of Bangladesh). Most subjects were 5-year-olds. One teacher in each school volunteered

to implement either Jolly Phonics (JP) or Big Books (BB). The JP and BB schools were well matched on most variables including free meals and academic performance, but the JP schools had many more children at beginning ESL levels (53% vs. 30%). Extensive batteries of pretests showed the groups to be very similar in measures of oral language, except that ESL students in BB scored significantly higher than those in JP on a vocabulary scale. There were few differences in premeasures of phoneme discrimination and letter knowledge. However, JP students had substantially higher pretests on measures of phoneme identification, segmentation, letter sound recognition, and sound writing (median ES = +0.88), and on reading and writing words (median ES=+0.70). As these were the same measures used as posttests, these pretest differences are of concern. The difference probably arose from the fact that the study began halfway through the school year, in January, and the teachers who chose JP may have already taught more phonics and reading than those who chose BB. Gain scores were used to deal with these pretest differences, but this is not a sufficient solution to pretest differences of this size.

The interventions took place one hour per day for 12 weeks. The results strongly favored the JP group. Effect sizes for five gain scores measures of phonemic awareness and phonics knowledge had a median value of +0.70 at posttest and +0.16 on a delayed posttest one year later. However, note that effect sizes for gain scores cannot be compared to those for point-in-time scores. On five measures of reading and writing, the median effect size for gain scores was +1.06 at the end of the experiment and +0.52 one year later.

Reading Recovery/Descubriendo la Lectura

Reading Recovery is an early intervention tutoring program for young readers who are experiencing difficulty in their first year of reading instruction (Clay, 1993). The program provides the lowest achieving readers (lowest 20%) in first grade with supplemental tutoring in addition to their regular reading classes. Children participating in Reading Recovery receive daily one-to-one 30-minute lessons for 12-20 weeks with a certified, specially trained teacher. The lessons include assessment, reading known stories, reading a story that was read once the day before, writing a story, working with a cut-up sentence, and reading a new book. Descubriendo la Lectura (DLL), the Spanish adaptation of Reading Recovery, is equivalent in all major aspects to the original program. There have been many evaluations comparing Reading Recovery and control students, including a large-scale, randomized evaluation in Ohio (Pinnell, Lyons, Deford, Bryk, & Seltzer, 1994). Only one study involving English language learners met the inclusion standards of this review. This was a 7-month evaluation of Descubriendo la Lectura (DLL) conducted by Escamilla (1994) in Tucson. The experiment compared 23 DLL students to 23 matched comparison students also taught in Spanish in another school. In both cases, students were identified as being in the lowest 20% of their classes based on individually administered tests and teacher judgment. The two groups were well matched on the Spanish Aprenda. The outcomes of DLL on Spanish reading measures at the end of first grade were very positive. On six scales of a Spanish Observation Survey adapted from the measures used in evaluations of the English Reading Recovery program, DLL students started out below controls and ended the year substantially ahead of them, with a median effect size of +0.84. These scores were also compared to those of a random sample of all students, most of whom were not having reading difficulties, and the

DLL students performed above the level of the classes as a whole on all scales. Students were also pre- and posttested on a standardized test, the Aprenda Spanish Achievement Test. On a total reading score, DLL students increased from NCEs of 37.7 to 45.2. Control students increased from NCEs of 36.5 to 37.7, while a comparison group of children not experiencing difficulties diminished from 41.9 to 39.5. Translating the experimental-control differences to normal curve equivalents and dividing by the theoretical standard deviation of NCEs (21.06), the effect size on the Aprenda was +0.30.

Small Group Tutorials with Direct Instruction

Gunn, Biglan, Smoklowski, and Ary (2000) evaluated a small group tutorial program that used two forms of DI, *Reading Mastery* and *Corrective Reading*, as a supplementary intervention for Hispanic and non-Hispanic children who were struggling in reading. The children were in kindergarten to third grade, and were selected either because they scored at a very low level on an achievement measure or because they were rated by their teachers as being high in aggressive behavior (and were below grade level in reading). Children were selected from nine rural Oregon elementary schools. They were randomly assigned to experimental or control conditions. Those children assigned to the experimental group were taught in homogeneous groups of one to three children using *Reading Mastery* if they were in grades K-2, or *Corrective Reading* if they were in grades 3-4. They were taught daily by instructional assistants for two years. Only 19 of the 122 Hispanic students were considered non-English speaking; the oral English skills of the remaining students were not specified.

The experimental and control groups were very well matched on the Woodcock-Johnson Letter Word Identification and Word Attack scales, and on Oral Reading Fluency. After the first year, tutorial students who had received five to six months of supplementary instruction showed greater gains than control students on all three measures, Letter-Word ID (ES=+0.22), Word Attack (ES=+0.70), and Fluency (ES=+0.16). Only the Word Attack differences were significant. At the end of the second year, after 15-16 months of instruction, effect sizes for gains from pretest on these measures were +0.46, +0.91, and +0.43, respectively. In addition, there were positive effects on Woodcock Reading Vocabulary (ES=+0.44) and Passage Comprehension (ES=+0.48), given as posttests only. Experimental-control differences on all five measures were significant after two years for all students (Hispanic and non-Hispanic), and there were no ethnicity x treatment interactions.

A special analysis was carried out for the 19 initially non-English speaking children, who did particularly well in the program compared to matched controls. Children in the experimental group made significantly greater gains only on mean words read per minute, but all other scores were in the same direction, though not significant given the small sample size.

Libros

Goldenberg (1990; see also Goldenberg, Reese, & Gallimore, 1992) studied a school and home reading intervention for Spanish-dominant kindergartners. The intervention, called Libros, involved teachers introducing and extensively discussing a Spanish story and then sending home photocopied "books" with children once every three weeks through kindergarten. Parents were encouraged to read

with their children and were shown a videotape of a parent reading and discussing the story. In control classrooms, teachers sent home worksheets on letters and syllables. Children in four classrooms using Libros were matched with those in four control classrooms based on Bilingual Syntax Measure scores. On an experimenter-constructed set of 13 Spanish early literacy assessments at the end of the year, experimental children scored significantly higher than controls (median ES=+0.51; MANCOVA for all 13 measures, $p<.001$). Effects were strongest on measures of letter and word identification, but less positive on comprehension measures. In a companion study, Goldenberg, Reese, and Gallimore (1992) observed five Libros and five control students at home, using their respective materials. To their surprise, however, parents used both sets of materials in similar ways, emphasizing copying and repetition rather than the relationship between print and meaning.

Studies of Upper Elementary Reading Programs

Several studies have evaluated reading programs for English language learners in grades 2-5. Ten of these met the inclusion criteria. These are summarized in Table 3 and described in the following sections.

TABLE 3
Upper Elementary and Secondary Reading Program: Descriptive Information and Effect Sizes for Qualifying Studies

Study	Intervention description	Design	Duration	N	Grade	Sample Characteristics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Upper Elementary										
Calderon et al (1998)	Bilingual Cooperative Integrated Reading & Composition (BCIRC)	Matched control	2 yrs	222	2-3	Spanish-dominant students in El Paso, TX	Well matched on demographics. Pretests results: 2nd grade cohort, E=C; 3rd grade cohort, C>E in Spanish (ES=-0.62) but E>C in English (ES=+0.26)	**Spanish TAAS**	**Grade 2**	Spanish Reading +0.30 Spanish Writing +0.62 English Reading +0.54 English Writing/ Language +0.29
								Reading	+0.30	
								Writing	+0.62	
								English TAAS	**Grade 3**	
								Reading	+0.54	
								Writing	+0.29	
								English TAAS	**2 yrs**	
								Reading	+0.87	
								Language	+0.38	
								English TAAS	**1 yr**	
								Reading	+0.33	
								Language	+0.22	
Saunders & Goldenberg (1996)	Enriched transition	Matched control	1 yr	140	2 & 5	Spanish-dominant students in Southern CA	Well matched on pretests	**English only group**		
								2nd grade-English Reading	+0.34	
								English Language	+0.42	
								5th grade-English Reading	+0.03	
								English Language	+0.72	
								TBE group		
								2nd grade-Spanish Reading	+1.36	
								Spanish Language	+1.37	
								5th grade-English Reading	+0.68	
								English Language	+0.81	
Saunders & Goldenberg (1999)	Enriched transition	Matched control	3 yrs	102	1-5	Spanish and Cantonese speaking students in Southern CA	Well matched on % of LEP, SES, ethnicity, and achievement scores	**Spanish subgroup**	**Spanish Measures**	
									Reading	Language
								1st grade	-0.02	+0.11
								2nd grade	+0.26	+0.20
								3rd grade	+0.38	+0.27
								4th grade	+0.59	+0.38
								Cantonese subgroup	**English measures**	
								4th grade	+0.53	+1.77
								5th grade	+0.80	+0.78
								English measures for early transition Spanish subgroup at 5th grade	+0.50	+0.56
								Spanish measures for late transition Spanish subgroup at 5th grade	+0.92	+0.69

Study	Intervention description	Design	Duration	N	Grade	Sample Characteristics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Carlo et al. (in press)	Direct instruction in key vocabulary	Matched control	2 yrs	~130	4 & 5	ELL students in CA, VA, and MA	Well matched on pretests	**Eng Vocab Assessment**		+0.21
								PPVT	-0.08	
								Polysemy prod	+0.33	
								Morphology	+0.22	
								Semantic Association	+0.21	
								Eng Reading Comp	+0.17	
Perez (1981)	Oral language activity	Matched control	3 mos	150	3	Mexican American ELL students in TX	Well matched on demographics and pretests, E>C, ES=+0.15	**English**	+0.97	+0.97
Denton (2000)	Read Well (Tutoring using systematic phonics)	Random assignment	4 mos	33	2-5	Spanish-dominant bilingual students in Texas	Well matched on WRMT pretests; E>C, ES=+0.32 (0.3<p<0.6)	**English--Read Well**		+0.51
								Word Identification	+0.55	
								Word Attack	+0.46	
								Passage Comprehension	+0.00	
								Fluency	+0.18	
								Accuracy	+0.78	
								Comprehension	+0.82	
	Read Naturally (Tutoring using repeated readings)			60			Well matched on WRMT pretests	**English--Read Naturally**		+0.08
								Word Identification	-0.05	
								Word Attack	-0.13	
								Passage Comprehension	+0.16	
								Fluency	+0.23	
								Accuracy	+0.30	
								Comprehension	+0.00	
Waxman (1994)	Effective Use of Time (EUOT); ESL in the Content Areas (ESLCA); Combination	Matched control	6 mos	325	1-5	Hispanic ELL students in a medium-sized metropolitan area in south central region US	Poorly matched on ITBS; EUOT>C; ES=+0.83; ESLCA>C; ES=+0.62	**English Reading**		
								EUOT	+0.37	+0.28 (EUOT)
								ESLCA	+0.01	
								Combined	-0.38	+0.04 (ESLCA)
								English Lang Arts		
								EUOT	+0.18	-0.33 (Combined)
								ESLCA	+0.07	
								Combined	-0.27	
Schon, Hopkins, & Davis (1982)	Literature in Spanish, free reading	Matched control	8 mos	39	2	Hispanic students in five 2nd grade classes in Tempe AZ	Poorly matched on pretests, C>E	**Spanish**		
								Reading Comprehension	E=C	Spanish E>C
								Reading vocabulary	E>C	
								English		
								Reading Comprehension	C>E	English E=C
								Reading vocabulary	E>C	

Study	Intervention description	Design	Duration	N	Grade	Sample Characteristics	Evidence of Initial Equality	Posttest	Effect Size	Median ES
Roser, Hoffman, & Forest (1990)	Literature units in English	Matched control	18 mo	5 E & 5 C classes	2	Spanish-dominant students in Brownsville TX	Marginal design, C>E, ES=-0.39 (NCE's) on CTBS reading and langage arts	**English CTBS**		
								Reading	+0.24	+0.22
								Language Arts	+0.19	
Hafiz & Tudor (1989)	Voluntary after school reading for pleasure	Matched control	3 mos	46	Age 10-11	Pakistani origin ELL students in London England	Well matched on demographics and fairly well on pretests, C>E	**English**		
								National Foundation for Educational Research Tests of Proficiency in English (only mean scores and t-ratio for gains)	E>C	
Secondary										
Schon, Hopkins, & Davis (1984)	High interest Spanish reading materials	Matched control	4 mos	111	9-12	Spanish dominant high school students in Tempe AZ	No pretest scores provided but pretest scores were used as covariates in the analysis	**Study I**		
								Spanish CTBS	-0.10	-0.10
								English MRT	-0.08	-0.08
			7 mos	40	10-12			**Study II**		
								Spanish CTBS	-0.11	-0.11
								English MRT	+0.09	+0.09
								English vocabulary	-0.08	-0.08
Schon, Hopkins, & Davis (1985)	High interest Spanish reading materials	Matched control	8.5 mos	190	7-8	Spanish dominant junior high school students in Tempe AZ	No pretest scores provided but pretest scores were used as covariates in the analysis	**Spanish Inter-American Series**	Across 3 measures	
								Grade 7	+0.28	+0.36
								Grade 8	+0.43	
								English Inter-American Series	Across 3 measures	
								Grade 7	-0.11	-0.14
								Grade 8	-0.16	
Shames (1998)	Community language learning model and comprehension processing	Semi-random	1 yr	58	9-12	Spanish and Haitain Creole speaking students in Palm Beach County, FL	C>E on pretests, ES=-0.59	**English Idea Proficiency Test--Reading**		
								Community language learning vs control	+0.46	+0.46
								Comprehension processing strategies vs control	+1.00	+1.00
								Combination vs control	+1.22	+1.22

Bilingual Cooperative Integrated Reading and Composition (BCIRC)

A large experiment by Calderón, Hertz-Lazarowitz, and Slavin (1998) evaluated a cooperative learning program called Bilingual Cooperative Integrated Reading and Composition, or BCIRC. BCIRC is an adaptation of Cooperative Integrated Reading and Composition, an upper elementary reading program based on principles of cooperative learning that has been successfully evaluated in several studies (see Stevens, Madden, Slavin, & Farnish, 1987). CIRC was the basis for the upper-elementary components of Success for All, described earlier. BCIRC was adapted to meet the needs of limited English proficient children in bilingual programs who are transitioning from Spanish to English reading. In CIRC and BCIRC, students work in four-member heterogeneous teams. After a teacher introduction, students engage in a set of activities related to a story they are reading. These include partner reading in pairs, and team activities focused on vocabulary, story grammar, summarization, reading comprehension, creative writing, and language arts. BCIRC adds to these activities transitional readers (in this study, Macmillan's *Campanitas de Oro* and *Transitional Reading Program*), and ESL strategies, such as total physical response, realia, and appropriate use of cognates, to help children transfer skills from Spanish to English reading.

Control teachers also used the same *Campanitas de Oro* and *Transitional Reading Program* textbooks, and received training in generic cooperative learning strategies. None of the control teachers used cooperative learning consistently, although all of them made occasional use of these strategies.

The BCIRC study involved 222 Hispanic children in the Ysleta Independent School District in El Paso, Texas. Seven of the highest-poverty schools in the district were assigned to experimental (3 schools) or control (4 schools) conditions. As a whole, the experimental and control schools were well matched demographically. Two cohorts were assessed, one of which was involved for just one year (second grade) and the other for two years (grades 2-3). End-of-kindergarten Bilingual Syntax Measure scores for both English and Spanish were nearly identical for experimental and control groups in the second-grade cohort, but in the third-grade cohort the experimental group scored significantly lower in Spanish (ES=-0.62) and somewhat higher in English (ES=+0.26). These BSM measures were used as covariates in all analyses.

Analyses of covariance found significantly higher scores for students in BCIRC classes in both cohorts. For the grade 2 cohort (one year of treatment), the effect size on the Spanish Texas Assessment of Academic Skills (TAAS) was +0.30 in reading ($p<.06$) and +0.54 in writing ($p<.02$). For the grade 3 cohort, the effect size for English TAAS reading was +0.62 ($p<.01$), and for English language, ES=+0.29 (n.s.). In the grade 2-3 cohort, half of the students actually experienced the treatment for two years and half for just one year (due to mobility and different patterns of class assignment). Analyzing these two groups separately, the two-year group had an effect size of +0.87 for reading and +0.38 for language, while the one-year subgroup had effect sizes of +0.33 for reading and +0.22 for language. Finally, in the grade 2-3 cohort, 32% of students met the state criterion for

exit from bilingual education in reading, compared to 10% in the control group ($p<.01$). In language, 39% of BCIRC and 21% of control students met the criterion ($p<.06$).

Many other studies of cooperative learning have found these strategies to increase student achievement when they incorporate group goals that can be attained only if all members individually achieve the academic objective (Slavin, 1995), as in BCIRC and in the cooperative learning methods in Success for All. Many of these involved schools serving Latino students, but only the BCIRC study focused on elementary reading (see Fashola, Slavin, Calderón, & Durán, 2001).

Enriched Transition

Saunders and Goldenberg (1996) evaluated a program designed to help English language learners transition from Spanish to English. The treatment focused on literature study, writing, discourse, skill building, reading comprehension strategies, independent reading, teacher read-alouds, and other elements. These treatments were applied to second and fifth graders in transitional bilingual education (TBE) and English-only classes. In each case, a control group was matched with the experimental group. Over a year, the English-only experimental group scored higher than control groups on an English reading measure in second grade (ES=+0.34), but not in fifth grade (ES=+0.03). Second grade TBE students, tested in Spanish, scored substantially better in the experimental condition (ES=+1.36). Fifth-grade experimental TBE students tested in English also showed substantially higher achievement (ES=+0.68).

The Saunders and Goldenberg (1996) article only reported on the first year of a three-year transition project. A study of the full program was described by Saunders (1998). It compared children in the three-year transition program (using the methods described above) to those in a three- to six-month transition, the usual treatment for ELLs in the district studied. On Spanish measures, differences were insignificant in grade 1 (ES= -0.02) and grade 2 (ES=+0.26), but significant in grade 3 (ES=+0.38). In a Cantonese-dominant subgroup, experimental students scored significantly higher on English tests (grade 4, ES=+0.53; grade 5, ES=+0.80). Experimental fourth graders assessed in Spanish also scored higher than controls (ES=+0.59). At fifth grade, an early-transitional group was tested in English and a late-transitioning group was tested in Spanish. In both cases, effects favored the experimental group (ES=+0.50 for English, ES=+0.92 for Spanish). Similar effects were seen on performance measures of reading and writing, and experimental students passed a test used as a criterion for placement in English-only instruction at much higher rates than did controls.

The Saunders (1998) and Saunders and Goldenberg (1996) studies provide strong evidence for enriched, carefully planned transition strategies to support students moving from either Spanish to English instruction or from ESL to unsupported English instruction. Although the samples were quite small, the experimental and control groups were well matched and showed clear positive effects of the transition treatment in the relevant grades, especially grades 3-5.

Vocabulary Intervention

Carlo, August, McLaughlin, Snow, Dressler, Lippman, Lively, and White (in press) carried out a two-year evaluation of a vocabulary teaching intervention with Spanish-dominant fourth and fifth graders in California, Massachusetts, and Virginia. The intervention involved introducing 12 vocabulary words each week using a variety of strategies, such as charades, 20 questions, discussions of Spanish cognates, word webs, and word association games.

The experimental students were taught in one five-week unit and two six-week units in the first year, and three five-week units in the second year. Matched control students continued their usual instruction. Experimental and control students were not significantly different on any of an extensive set of measures.

At the end of the first year, ELLs showed greater gains from pretest than controls (ES= +0.27), but surprisingly, gains were lower after two years of intervention (ES is not given, but is estimated at +0.17). Gains on the Peabody Picture Vocabulary Test were not significant (ES estimated at –0.08). However, assessments of skills closer to those that were taught showed larger gains for experimental than for control students. This includes a test of the specific words taught in the experimental treatment, but more importantly, experimental students after two years of intervention gained more on a test of polysemy (the ability to generate multiple meanings of words), ES=+0.22, a test of morphology (the ability to use proper forms of words), ES=+0.18, and a test of semantic associations (e.g., knowing that the words "dog" and "barks" are associated), ES=+0.23.

Carlo et al. (in press) also evaluated the treatments with English-only students, who showed a very similar pattern of gains (although they began and ended the experiment with significantly higher scores than the ELLs). This finding suggests that the treatments were generally effective in promoting children's vocabulary development, rather than being uniquely appropriate for ELLs.

In a related study, Perez (1981) evaluated an oral language intervention with Spanish-dominant third graders in Texas. The intervention consisted of daily 20-minute sessions in which children worked with humorous language games, pictures, and other activities intended to build their oral abilities. The experimental group of 75 students was compared to a well-matched control group. On an unspecified reading measure, the experimental group scored substantially higher (ES=+0.97). If the posttest was not specific to the treatment, this would be an important finding, but the nature of the posttest is unclear.

Tutoring

Two types of one-to-one tutoring for English language learners were studied in a dissertation by Denton (2000). Spanish-dominant students in grades 2-5 in a bilingual program in Texas were assigned to one of two separate experiments. Those scoring lower than the first-grade level on the Woodcock Word Attack scale were randomly assigned to a program called "Read Well" (Sprick, Howard, & Fidanque, 1998), or to an untutored control group. Those scoring higher than this were randomly assigned to a tutoring program called "Read Naturally" or to an untutored control group. Read Well uses systematic phonics instruction and practice in fully decodable text (like the first-

grade instruction in Success for All). Read Naturally (Ihnot, 1992) emphasized repeated readings of connected text, vocabulary, and comprehension instruction. Tutors were undergraduate education majors. All tutoring was done in English. The final sample of students in the Read Well evaluation included 19 experimental and 14 control children. Those in the experimental group received an average of 22 tutoring sessions. In the Read Naturally comparison, there were 32 tutored and 28 non-tutored children.

The results indicated substantially higher achievement for the Read Well students than for controls, with a median effect size of +0.51 across six measures. Differences were statistically significant only on the Woodcock Word Attack scale ($p<.016$) and an oral reading accuracy scale ($p<.001$). In contrast, there were no differences between the children tutored with Read Naturally and those who were not tutored (ES= +0.08).

Effective Use of Time and ESL in the Content Areas

Waxman, Walker de Felix, Martínez, Knight, and Padrón (1994) evaluated two programs and a combination to improve the reading performance of English language learners in grades 1-5. They did not separate results by grade level, but as most students were in grades 2-5, this study is included in this section.

One of the programs, ESL in the Content Areas (ESLCA; Chamot & O'Malley, 1987), was designed to foster English language development by explaining concepts in Spanish and then using graphic mapping and problem-solving activities to help children transfer their understandings to English. The other, Effective Use of Time (EUOT; Stallings, 1986) consisted of a lesson structure for English-language instruction that emphasized pretesting, informing, guided practice, and posttesting.

Teachers of grades 1-5 in five schools were recruited for the experiment. Two schools were assigned to the control group, and then one each was assigned to EUOT, ESLCA, or a combined intervention. There were 17 bilingual teachers and 325 Hispanic English language learners in the study across the five schools. Pretest differences among the four groups were substantial, unfortunately. The EUOT group started out with Iowa Test of Basic Skills reading scores that were 83% of a standard deviation above controls, and the ESLCA group was 62% of a standard deviation ahead at pretest. Analyses of covariance found an overall effect of treatments for both reading ($p<.001$) and language ($p<.05$). Pairwise comparisons were not carried out, but the EUOT students scored highest on both reading (ES=+0.37) and language (ES=+0.18), and the combined intervention group scored below the control group (ES=-0.38 and -0.27, respectively). The ESL in the Content Areas group was almost identical to the control group on both measures.

Due to the pretest differences and the use of undifferentiated reading assessments across grades 1-5, this study is far from conclusive with respect to beginning reading. For reading in general, it counsels caution in using ESL in the Content Areas to improve the English reading achievement of English language learners, and may support the use of the Effective Use of Time lesson structure.

Literature Programs

Three studies investigated the use of various interventions focused on extensive use of children's literature with English language learners.

Schon, Hopkins, and Davis (1982) carried out a small, eight-month experiment in which about 37 Hispanic children in five Spanish bilingual, second-grade classes were non-randomly assigned to one of two conditions. Experimental students were given an extensive library of books in Spanish, and teachers were asked to provide at least 60 minutes per week of free reading time. Control teachers primarily taught reading in English, and did not receive additional books or training.

Unfortunately, the experimental and control groups were quite different at pretest, with control groups scoring much higher. On posttests adjusted for pretests, students in the experimental group scored significantly better on a measure of Spanish reading vocabulary ($p < .05$), but not Spanish reading comprehension. On English reading vocabulary the experimental group scored marginally better than controls ($p < .07$), but controls marginally outscored the experimental group in English reading comprehension ($p < .10$). The article reports effect sizes, but does not report standard deviations, and appears to have used adjusted, rather than raw, standard deviation units, which greatly inflate estimates of effect sizes. Therefore, effect size estimates are not shown in Table 2.

Roser, Hoffman, and Farest (1990) evaluated a program in which Hispanic students in Brownsville, Texas, were given extensive libraries of children's literature, accompanied by specially written units designed to develop themes, build vocabulary, encourage read-alouds and writing, and so on. All instruction was apparently in English. Most students were Hispanic, but no data showed how many were Hispanic or how many were English language learners in experimental or comparison groups. Treatments were implemented over 18 months, when students were in grades K-2.

The data presented were CTBS scores for second graders, collected as part of the district's usual accountability system. No pretests were available. Two kinds of comparisons were made. All second graders in the six experimental schools were compared to those in the same school the previous year. On average, there was a Normal Curve Equivalent (NCE) gain of 6.4 points in reading and 7.8 in language arts. In addition, six "comparison" schools were designated (although their characteristics were not described). These schools were not well matched with the experimental group, as the cohort before the study began scored 7.3 NCEs ahead of the previous cohort in the experimental schools. The cohort gain in these comparison schools averaged 1.4 NCEs in reading and 3.7 in language arts. No statistical tests were used. Using the theoretical standard deviation for normal curve equivalents (21.06), the differential gains yielded approximate effect sizes of +0.24 for reading and +0.19 for language arts, but these estimates should be interpreted with great caution.

In a small study with Panjabi-dominant children in Leeds, England, Hafiz and Tudor (1989) evaluated a program in which children ages 10-11 volunteered for an after-school reading program at a mosque. For one hour each afternoon, they selected graded readers, which they read on site or at home, and discussed books with the other students. On a series of nationally normed tests, the 16 children in the experimental group gained significantly over the 12-week period. Control students in the same school and in a control school with similar pretests showed no significant gains. Effect sizes could not be computed. Because the experimental students were self-selected, control students did

not have additional reading time, and the number of students was very small, this study must be interpreted with caution.

Secondary Reading

Only four secondary studies qualified for this review. Three of these were by the same group of authors. These are summarized in Table 3. Two studies were reported by Schon, Hopkins, and Vojir (1984). The first took place in a high-income high school in Tempe, Arizona, to which many low-income Hispanic students were bused. Low-achieving Hispanic students in grades 9-12 were assigned to experimental or matched control classes for four months. The treatment involved introducing high-interest Spanish newspapers, magazines, and books into remedial reading classes otherwise taught in English. Teachers were asked to give students 45 minutes each week to look at these materials on their own. Posttests found no differences in English reading (ES=-0.08) or in Spanish reading (ES=-0.10).

A second study at a similar high school used a similar treatment and design over seven months. Again, there were no differences in English reading (ES=-0.11) or in Spanish reading (ES=+0.09).

Finally, Schon, Hopkins, and Vojir (1985) compared seventh and eighth graders in a Tempe junior high school to similar students in the previous year. The treatments were the same as in the high school studies, but were implemented over a full school year. Averaging across three reading measures in each language, control groups scored better (adjusting for pretests) in English in both seventh grade (ES= -0.11) and eighth grade (ES= -0.16), although only the seventh-grade reading comprehension difference was statistically significant. On Spanish measures, the experimental group scored higher in both seventh (ES=+0.28) and eighth grades (ES=+0.43).

A study by Shames (1998) with Haitian Creole and Spanish-speaking students in South Florida evaluated three treatments. One, a "community language learning" model, involved students in grades 9-12 who were in an English as a Second Language class working in cooperative groups to write, record, and discuss their own English stories in addition to more usual commercially published texts. A second group, "comprehension processing," focused on graphic organizers, question-answer strategies, and K-W-L activities in which students discussed what they knew, wanted to know, and then learned about given topics. They read the stories generated by the "community language learning" group as well as commercial texts. The third treatment combined the first two, and there was a traditional control group. Students were assigned to treatments in a semi-random strategy, in which matched classes were assigned at random to treatment conditions. The treatments were implemented over a full school year.

Results indicated that taken together, students in the three treatments scored higher than controls, controlling for pretests, on a standardized Idea Proficiency Test reading scale (ES=+0.76). The combined treatment scored highest (ES=+1.22), followed by the comprehension processing treatment (ES=+1.00) and the community learning intervention (ES=+0.46), with the control group scoring lowest. Only the overall comparison was statistically significant, however.

Conclusions: Studies of Reading

The research summarized in this report shows how much remains to be done on effective reading programs for English language learners. Only a handful of studies met the minimal inclusion standards applied in this review, which principally required an experimental-control comparison of a reading program over at least 12 weeks, with evidence that the two groups were equivalent at pretest.

Beginning Reading. Among the 11 studies of interventions beginning in kindergarten or first grade that met these standards, the largest number involved Success for All, a comprehensive reform model (Slavin & Madden, 1999). Two studies of Success for All in its Spanish bilingual form found consistent, though highly variable, positive effects on students' Spanish reading performance (in comparison to schools teaching in Spanish using alternative methods). Effect sizes for first graders averaged +0.22 in one study (Nunnery et al., 1997), and +0.97 in another (Livingston & Flaherty, 1997). Similarly, schools using the English language adaptation of Success for All with Latino and Asian English language learners found positive but, again, highly variable effects, ranging from effect sizes of +0.24 to +1.36 for first graders. Longer-term studies through grades 3-5 found even more diverse outcomes, ranging from non-significant differences at third grade (Livingston & Flaherty, 1997) due to transitioning of the most successful students to English-only classes, to effect sizes in excess of +1.0 in grades 4-5 (Slavin & Madden, 1995). A four-year Texas study (Hurley et al., 2001) found that Hispanics showed greater gains on state reading assessments in Success for All schools than in control schools. While these studies generally support the effectiveness of Success for All for ELL and language minority children, the great variability in the outcomes suggests that more research is needed to understand these effects.

Two longitudinal studies found strong and lasting effects of Direct Instruction (DI) on the reading achievement of language minority students. One was a followup of mostly Hispanic fifth and sixth graders in Texas who had experienced DI in grades K-3 (Becker & Gersten, 1982). The other was a two-year study of DI in a structured immersion program for Asian English language learners (Gersten, 1985). An adaptation of DI for use in small-group tutorials (1-3 children) also found positive effects (Gunn et al, 2000).

No other beginning reading program had more than a single methodologically adequate study. A study of a systematic phonics program called Jolly Phonics (Stuart, 1999) found promising effects among children of Bangladeshi origin in London, but the study had serious problems with pretest differences. Very positive effects were documented in a study of a Spanish adaptation of Reading Recovery (Escamilla, 1994). A study of Libros, a home and school literature approach using Spanish reading materials, documented benefits for ELL kindergartners (Goldenberg, 1990).

Upper Elementary Reading. Ten studies of reading in grades 2-5 met the inclusion criteria. A two-year evaluation of Bilingual Cooperative Integrated Reading and Composition (BCIRC; Calderón et al., 1998), a cooperative learning strategy, found strong positive effects on the Spanish and English reading of children transitioning from Spanish to English reading in grades 2-3. Saunders (1998) and Saunders and Goldenberg (1999) successfully evaluated an enriched transition process for ELLs moving to English-only instruction. Carlo et al. (in press) found positive effects of an English vocabulary intervention for ELL fourth and fifth graders on various experimenter-made measures of vocabulary skill, and Perez (1981) found that instruction in oral English skills improved the reading

skills of ELL third graders. Denton (2000) evaluated two tutoring approaches and found that Read Well, a phonetic program, improved the English reading of very low achieving ELLs. Other studes found promising effects of programs emphasizing literature (Schon et al., 1982; Roser et al., 1990; Hafiz & Tudor, 1989) and Effective Use of Time (Waxman et al., 1994).

Secondary Reading. Only one of the four secondary studies that met the inclusion criteria found significant positive effects. Shames (1998) evaluated three models that made extensive use of cooperative learning and direct instruction in comprehension strategies. All three methods helped low-achieving speakers of Haitian Creole and Spanish accelerate their reading achievement more than members of a control group receiving traditional instruction.

The evidence cited here is consistent with the conclusion reached by Fitzgerald (1995) that effective beginning reading programs for English language learners are likely to be similar to those for English proficient children, with appropriate adaptations to their language proficiency. The programs with the strongest evidence of effectiveness in this review are all programs that have also been found to be effective with students in general: Success for All (Slavin & Madden, 2000, 2001), Direct Instruction (Adams & Engelmann, 1996); Reading Recovery (Pinnell et al, 1994), and phonetic tutoring (e.g., Wasik & Slavin, 1993). In fact, several of the studies evaluating Success for All (e.g., Nunnery et al. 1997; Livingston & Flaherty, 1997; Ross et al., 1998) as well as DI (Gunn et al., 2000), also included non-ELL students, and in each case those students also gained from the interventions, to about the same degree. The beginning reading programs with the strongest evidence of effectiveness in this review made use of systematic phonics, such as Success for All, Direct Instruction, and Jolly Phonics, but systematic phonics has been identified as a component of effective beginning reading programs for English proficient students as well (see National Reading Panel, 2000; Gersten & Geva, 2003). It may be that programs that are quite different from these exist but have not been adequately evaluated, or could be developed. The observation, however, that currently available reading methods known to be effective for English proficient students also accelerate the achievement of English language learners implies that a broader range of interventions also known to be effective with children in general may likewise be effective with English language learners, with appropriate adaptations (see Slavin & Calderón, 2001).

OVERALL CONCLUSIONS

While there is much more we need to know about reading instruction for English language learners, existing research does provide some empirically supported suggestions for policy and practice. First, there is a good deal of support for the idea that native language instruction can be beneficial for the English reading of English language learners. Not every study finds this to be true, but the higher-quality, longitudinal studies involving treatments of at least three years support this practice, including all three randomized studies in elementary schools and one of the two randomized secondary studies. None of the studies that met the inclusion standards found bilingual education to be significantly worse than immersion in enhancing English reading outcomes.

A surprising finding, however, is that many of the studies showing positive effects of bilingual education use paired bilingual strategies that teach reading in English and in the native language at the same time (at different times of the day), or that use a very fast transition (e.g., one year in Spanish before beginning transition). It may be that existing methodologically adequate studies have not followed children long enough to adequately evaluate bilingual programs that delay English instruction to third grade or later, but we did not find any convincing evidence to support the idea that English language learners need to wait before beginning English reading instruction, if they are also receiving reading instruction in the native language in the early years.

Teaching reading in two languages, with appropriate adaptations of the English program for the needs of English language learners, may represent a satisfactory resolution to the acrimonious debates about bilingual education. Proponents of bilingual education want to launch English language learners with success while maintaining and valuing the language they speak at home. Opponents are concerned not so much about the use of native language, but about delaying the use of English. Paired bilingual models immerse children in both English reading and native language reading at the same time. They are essentially half of a two-way bilingual model; by encouraging English proficient students to also take Spanish reading, any school with a paired bilingual model can readily become a two-way program, offering English-only children a path to early acquisition of a valuable second language.

Language of instruction must be seen as only one aspect, however, of instructional programming for English language learners. As many previous reviewers have concluded, quality of instruction is at least as important as language of instruction. This synthesis identified a number of specific, replicable programs that have strong evidence of effectiveness. Particularly well supported are Success for All and Direct Instruction, two well-structured, phonetic reading approaches that provide appropriate English language development adaptations for ELLs. Success for All also offers a Spanish version for use in bilingual models, which has been successfully evaluated. A British study of a program called Jolly Phonics similarly supports systematic phonics with English language learners in the early grades. Further, a study of Bilingual Cooperative Reading and Composition (BCIRC) supports the value of combining cooperative learning and cognitive strategy instruction, especially in helping children actively use English as they transition from Spanish to English reading. A high school study also found benefits of combining cooperative learning and cognitive strategy instruction (Shames, 1998). Beyond the use of systematic phonics and cooperative learning, there is evidence

from two studies to support the effectiveness of one-to-one tutoring for ELLs who are struggling in reading, and Gunn et al. (2000) showed positive effects of using DI in small groups of children. There is evidence that direct teaching of English vocabulary can help the reading performance of ELLs. Finally, a few studies found that encouraging children to read a wide range of grade-appropriate books helps to build their reading skills.

Clearly, language of instruction and replicable models are not mutually exclusive issues. Effective reading models can be applied in English, in the native language, or in both languages.

While we do have a good start on research in several areas, there is much more to be done. Large-scale, randomized, longitudinal evaluations of well-justified approaches are needed to more confidently recommend effective strategies for English language learners of all ages and backgrounds. Research systematically varying program components and research combining quantitative and qualitative methods are needed to more fully understand how various interventions affect the development of reading skills among English language learners. It is time to end the ideological debates, and to instead focus on good science, good practice, and sensible policies for children whose success in school means so much to themselves, their families, and our nation's future.

REFERENCES

Adams, G.L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

Alvarez, J. (1975). *Comparison of academic aspirations and achievement in bilingual versus monolingual classrooms.* Doctoral dissertation. UT Austin.

Ames, J., & Bicks, P. (1978). *An evaluation of Title VII bilingual/bicultural program, 1977-1978 school year, final report. Community School District 22.* Brooklyn, NY: School District of New York.

Ariza, M. (1988, April). *Evaluating limited English proficient students' achievement: Does curriculum content in the home language make a difference?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

August, D. (2002). *English as a second language instruction: Best practices to support the development of literacy for English language learners*. Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk.

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

Bacon, H., Kidd, G., & Seaberg, J. (1982). The effectiveness of bilingual instruction with Cherokee Indian students. *Journal of American Indian Education, 21* (2), 34-43.

Baker, K. (1987). A meta-analysis of selected studies in the effectiveness of bilingual education. *Review of Educational Research, 57*, 351-362.

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.

Baker, K., & de Kanter, A. (1983). An answer from research on bilingual education. *American Education, 56*, 157-169.

Baker, S., & Gersten, R. (1997). *Exploratory meta-analysis of instructional practices for English-language learners*. (Tech Rep. No. 97-01). Eugene, OR: Eugene Research Institute.

Balasubramonian, K., Seelye, H., & Elizondo de Weffer, R. (1973). *Do bilingual education programs inhibit English language achievement: A report on an Illinois experiment*. Paper presented at the 7th Annual Convention of Teachers of English to Speakers of Other Languages. San Juan.

Barclay, L. (1969). *The comparative efficacies of Spanish, English, and bilingual cognitive verbal instruction with Mexican American Head Start children*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Barik, H, & Swain, M. (1975). Three year evaluation of a large-scale early grade French immersion program: The Ottawa study. *Language Learning*. Vol. 25. No. 1. pp. 1-30.

Barik, H., & Swain, M. (1978). *Evaluation of a bilingual education program in Canada: The Elgin study through grade six*. Commission Interuniversitaire Suisse de Linguistique Appliquee. Switzerland.

Barik, H., Swain, M., & Nwanunobi, E.A. (1977). English-French bilingual education: The Elgin study through grade five. *Canadian Modern Language Review.* 33, 459-475.

Bates, E., & May, B. (1970). *The effects of one experimental bilingual program on verbal ability and vocabulary of first grade pupils*. Doctoral dissertation, Texas Tech University.

Becker, W.C., & Gersten, R. (1982). A follow-up on Follow Through: The later effects of the Direct Instruction model on children in fifth and sixth grades. *American Educational Research Journal, 19* (1), 75-92.

Brisk, M.E. (1998). *Bilingual education*. Mahwah, NJ: Erlbaum.

Bruck, M., Lambert, W.E., & Tucker, G.R. (1977). Cognitive consequences of bilingual schooling: The St. Lambert Project Through Grade Six. *Linguistics.* (24), 13-33.

Burkheimer, G.J., Conger, A.J., Dunteman, G.H., Elliott, B.G., & Mowbray, K.A. (1989). *Effectiveness of services for language-minority limited-English-proficient students.* Washington, DC: U.S. Department of Education.

Calderón, M. (2001). Curricula and methodologies used to teach Spanish-speaking limited English proficient students to read English. In R.E. Slavin & M. Calderón (Eds.), *Effective programs for Latino students*. Mahwah, NJ: Erlbaum.

Calderón, M., Hertz-Lazarowitz, R., & Slavin, R.E. (1998). Effects of Bilingual Cooperative Integrated Reading and Composition on students making the transition from Spanish to English reading. *Elementary School Journal, 99,* (2), 153-165.

Calderón, M., & Minaya-Rowe, L. (2003). *Designing and implementing two-way bilingual programs.* Thousand Oaks, CA: Corwin.

Carlisle, J.F., & Beeman, F.F. (2000). The effects of language of instruction on the reading and writing achievement of first-grade Hispanic children. *Review of Educational Research, 55* (3), 269-317.

Campeau, P.L., Roberts, A. Oscar H., Bowers, J.E., Austin, M., & Roberts, S.J. 1975. *The identification and description of exemplary bilingual education programs.* Palo Alto, CA: American Institutes for Research.

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

Carsrud, K., & Curtis, J. (1979). *ESEA Title VII bilingual program: Final report*. Austin, TX: Austin Independent School District.

Carsrud, K., & Curtis, J. (1980). *ESEA Title VII bilingual program: Final report.* Austin Independent School District. Austin.

Chamot, A., & O'Malley, J. (1987). The cognitive academic language learning approach: A bridge to the mainstream. *TESOL Quarterly, 21*, 227-249.

Christian, D., & Genesee, F. (Eds.). (2001). *Bilingual education.* Alexandria, VA: TESOL.

Ciriza, F. (1990). *Evaluation report of the preschool project for Spanish-speaking children, 1989-90.* San Diego, CA: San Diego City Schools.

Clay, M.M. (1993). *Reading Recovery: A guidebook for teachers in training*. Portsmouth, NH: Heinemann.

Cohen, A.D. (1975). *A sociolinguistic approach to bilingual education.* Rowley, MA: Newbury House Press.

Cohen, A.D., Fathman, A.K., & Merino, B. (1976). *The Redwood City Bilingual Education Report, 1971-1974: Spanish and English proficiency, mathematics, and language-use over time.* Toronto: Ontario Institute for Studies in Education.

Cooper, H. (1998). *Synthesizing research* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H.M., & Hedges, L.V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cottrell, M.C. (1971, April). *Bilingual education in San Juan Co., Utah: A cross-cultural emphasis.* Paper presented at the annual meeting of the American Educational Research Association, New York City.

Covey, D.D. (1973). *An analytical study of secondary freshmen bilingual education and its effects on academic achievement and attitudes of Mexican American students.* Doctoral dissertation, Arizona State University.

Curiel, H. (1979). *A comparative study investigating achieved reading level, self-esteem, and achieved grade point average given varying participation*. Doctoral dissertation. Texas A&M University.

Curiel, H., Stenning, W., & Cooper-Stenning, P. (1980). Achieved ready level, self-esteem, and grades as related to length of exposure to bilingual education. *Hispanic Journal of Behavioral Sciences,* Vol. 2, pp. 389-400.

Danoff, M.N., Arias, B.M., Coles, G.J., & others. (1977a). *Evaluation of the impact of ESEA Title VII Spanish/English bilingual education programs.* Palo Alto, CA*:* American Institutes for Research.

Danoff, M.N., Coles, G.J., McLaughlin, D.H. & Reynolds, D.J. (1978). *Evaluation of the impact of ESEA Title VII Spanish/English bilingual education programs, Vol. III: Year two impact designs.* Palo Alto, CA*:* American Institutes for Research.

Danoff, M. N., Coles, G. J., McLaughlin, D.H., & Reynolds, D. J. (1977b). *Evaluation of the Impact of ESEA Title VII Spanish/English bilingual education programs, Vol. I: Study design and interim findings.* Palo Alto, CA*:* American Institutes for Research.

Day, E.M., & Shapson, S.M. (1988). *Provincial assessment of early and late French immersion programs in British Columbia, Canada*. Paper presented at the annual meeting of the American Educational Research Associates. New Orleans.

de la Garza, V.J., & Marcella, M. (1985). Academic achievement as influenced by bilingual Instruction for Spanish-dominant Mexican American children. *Hispanic Journal of Behavioral Sciences. (7),* 3, 247-259.

de Weffer, R. (1972). *Effects of first language instruction in academic and psychological development of bilingual children.* Doctoral dissertation, Illinois Institute of Technology.

Denton, C.A. (2000). *The efficacy of two English reading interventions in a bilingual education program*. Unpublished doctoral dissertation, Texas A&M University.

Dianda, M., & Flaherty, J. (1995, April). *Effects of Success for All on the reading achievement of first graders in California bilingual programs*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Doebler, L.K. & Mardis, L.J. (1980-81). Effects of a bilingual education program for Native American children. *NABE Journal, 5* (2), 23-28.

Educational Operations Concepts, Inc. (1991a). *An evaluation of the Title VII ESEA bilingual education program for Hmong and Cambodian students in junior and senior high school.* St. Paul, MN: Educational Operations Concepts.

Educational Operations Concepts, Inc. (1991b). *An evaluation of the Title VII ESEA bilingual education program for Hmong and Cambodian students in kindergarten and first grade* St. Paul, MN: Educational Operations Concepts.

El Paso Independent School District. (1987). *Interim report of the five-year bilingual education pilot 1986-1987 school year.* El Paso, TX: Office for Research and Evaluation.

El Paso Independent School District. (1990). *Bilingual education evaluation: The sixth year in a longitudinal study*. El Paso, TX: Office for Research and Evaluation.

El Paso Independent School District. (1992). *Bilingual education evaluation*. El Paso, TX: Office for Research and Evaluation.

Escamilla, K. (1994). Descubriendo la Lectura: An early intervention literacy program in Spanish. *Literacy, Teaching, and Learning, 1* (1), 57-70.

Fashola, O.S., Slavin, R.E., Calderón, M., & Durán, R. (2001). Effective programs for Latino students in elementary and middle schools. In R.E. Slavin & M. Calderón (Eds.), *Effective programs for Latino students*. Mahwah, NJ: Erlbaum.

Fitzgerald, J. (1995). English as a second language instruction in the United State: A research review. *Journal of Reading Behavior, 27,* 115-152.

Garcia, G. (2000). Bilingual children's reading. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of reading research*, Vol. III (pp. 813-834). Mahwah, NJ: Erlbaum.

Genesee, F., Holobow, N. E., Lambert, W. E, & Chartrand, L. (1989). Three elementary school alternatives for learning through a second language. *The Modern Language Journal, 73*. 250-263.

Genesee, F., Lambert, W.E., & Tucker, G.R. (1977). *An experiment in trilingual education.* Montreal, Canada: McGill University.

Genesee, F. & Lambert, W.E. (1983). Trilingual education for majority-language children. *Child Development, 54,* 105-114.

Genesee, F., Lambert, W.E., Sheiner, E., & Tucker, G.R. (1983). *An experiment in trilingual education.* Montreal, Canada: McGill University.

Gersten, R. (1985). Structured immersion for language minority students: Results of a longitudinal evaluation. *Educational Evaluation and Policy Analysis, 7* (3), 187-196.

Gersten, R., & Geva, E. (2003). Teaching reading to early language learners. *Educational Leadership, 60* (8), 44-49.

Gersten, R., & Woodward, J. (1995). A longitudinal study of transitional and immersion bilingual education programs in one district. *The Elementary School Journal, 95* (3), 223-239.

Goldenberg, C. (1990, April). *Evaluation of a balanced approach to literacy instruction for Spanish-speaking kindergartners*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Goldenberg, C. (1996). The education of language-minority students: Where we are, and where do we need to go? *Elementary School Journal, 36* (4), 715-738.

Goldenberg, C., Reese, L., & Gallimore, R. (1992). Effects of literacy materials from school on Latino children's home experiences and early reading achievement. *American Journal of Education, 100* (4), 497-536.

Greene, J.P. (1997). A meta-analysis of the Rossell and Baker review of bilingual education research. *Bilingual Research Journal, 21* (2/3).

Grigg, W., Daane, M., Jin, Y., & Campbell, J. (2003). *The nation's report card: Reading 2002.* Washington, DC: US Department of Education.

Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education, 34* (2), 90-103.

Hakuta, K., Butler, Y.G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute, Policy Report 2000-1.

Hafiz, F.M., & Tudor, I. (1989). Extensive reading and the development of language skills. *ELT Journal,* 34(1), 5-13.

Howard, E.R., Sugarman, J., & Christian, D. (2003). *Two-way immersion education: What we know and what we need to know.* Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk.

Hurley, E.A., Chamberlain, A., Slavin, R.E., & Madden, N.A. (2001). Effects of Success for All on TAAS Reading: A Texas statewide evaluation. *Phi Delta Kappan*, 82 (10), 750-756.

Huzar, H. (1973). *The effects of an English-Spanish primary grade reading program on second and third grade students*. M.Ed. thesis, Rutgers University.

Ihnot, C. (1992). *Read naturally*. St. Paul, MN: Read Naturally.

Kaufman, M. (1968). Will instruction in reading Spanish affect ability in reading English? *Journal of Reading, 11*, 521-527.

Lambert, W.E., & Tucker, G.R. (1972). *Bilingual education of children: The St. Lambert Experience.* Rowley, Quebec: Newbury House.

Lampman, H.P. (1973). *Southeastern New Mexico Bilingual Program: Final report.* Artesia, NM: Artesia Public Schools.

Layden, R. G. (1972). *The relationship between the language of instruction and the development of self-concept, classroom climate, and achievement of Spanish speaking Puerto Rican children.* Doctoral dissertation, University of Maryland.

Legarreta, D. (1979). The effects of program models on language acquisition by Spanish-speaking children. *TESOL Quarterly, 13* (4), 521-534.

Livingston, M., & Flaherty, J. (1997). *Effects of Success for All on reading achievement in California schools*. Los Alamitos, CA: WestEd.

Lum, J.B. (1971). *An effectiveness study of English as a second language (ESL) and Chinese bilingual methods*. Doctoral dissertation, University of California at Berkeley.

Maldonado, J.A. (1994). Bilingual special education: Specific learning disabilities in language and reading. *Journal of Education Issues of Language Minority Students, 14,* 127-147.

Maldonado, J.R. (1974). *The effect of the ESEA Title VII program on the cognitive development of Mexican American students*. Doctoral dissertation, University of Houston.

Maldonado, J.R. (1977). *The effect of the ESEA Title VII program on the cognitive development of Mexican American students*. Unpublished doctoral dissertation, University of Houston, Houston, TX.

Malherbe, E.C. (1946). *The bilingual school.* London, England: Longmans Green.

Matthews, T. (1979). *An investigation of the effects of background characteristics and special language services on the reading achievement and English fluency of bilingual students*. Seattle, WA: Seattle Public Schools: Department of Planning, Research, and Evaluation.

McConnell, B.B. (1980a). *Effectiveness of individualized bilingual instruction for migrant students.* Doctoral dissertation, Washington State University

McConnell, B.B. (1980b). *Individualized bilingual instruction, Final evaluation, 1978-1979 program.* Pullman, WA: Washington State University.

McSpadden, J. (1979). *Acadiana bilingual bicultural education program: Interim evaluation report, 1978-79*. Lafayette Parish, LA.

McSpadden, J. (1980*). Acadiana bilingual bicultural education program: Interim evaluation report, 1979-80*. Lafayette Parish, LA.

Medina, M., & Escamilla, K. (1992). Evaluation of transitional and maintenance bilingual programs. *Urban Education,* 27(3), 263-290.

Melendez, W. A. (1980). *The effect of the language of instruction on the reading achievement of limited English speakers in secondary schools.* Doctoral dissertation. Loyola University of Chicago.

Meyer, M.M. & Fienberg, S.E. (1992). *Assessing evaluation studies: The case of bilingual education strategies.* Washington, DC: National Academy of Sciences.

Moore, F.B. & Parr, G.D. (1978). Models of bilingual education: Comparisons of effectiveness. *The Elementary School Journal. 79,* 93-97.

Morgan, J.C. (1971). *The effects of bilingual instruction of the English language arts achievement of first grade children.* Doctoral dissertation, Northwestern State University of Louisiana.

Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Rockville, MD: National Institute of Child Health and Human Development.

Nunnery, J., Slavin, R., Ross, S., Smith, L., Hunter, P., & Stubbs, J. (1997, March). *Effects of full and partial implementations of Success for All on student reading achievement in English and Spanish.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Olesini, J. (1971). *The effect of bilingual instruction on the achievement of elementary pupils*. Unpublished doctoral dissertation, East Texas State University, Commerce, TX.

Pena-Hughes, E, & Solis, J. (1980). *AB's.* McAllen, TX: McAllen Independent School District.

Perez, E. (1981). Oral language competence improves reading skills of Mexican-American third graders. *The Reading Teacher*, October, 24-27.

Pinnell, G., DeFord, D., & Lyons, C. (1988). Reading Recovery: Early intervention for at-risk first graders. Arlington, VA: Educational Research Service.

Pinnell, G., Lyons, C.A., DeFord, D.E., Bryk, A.S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high risk first graders. *Reading Research Quarterly, 29,* 9-40.

Plante, A.J. (1976). *A study of effectiveness of the Connecticut "Pairing" model of bilingual/bicultural education.* Hamden, CT: Connecticut Staff Development Cooperative.

Powers, S. (1978). *The influence of bilingual instruction on academic achievement and self-esteem of selected Mexican American junior high school students.* Doctoral dissertation, University of Arizona.

Prewitt Diaz, J.O. (1979). *An analysis of the effects of a bicultural curriculum on monolingual Spanish ninth graders as compared with monolingual English and bilingual ninth graders with regard to language development, attitude toward school, and self-concept*. Doctoral dissertation, University of Connecticut.

Ramirez, J., Pasta, D.J, Yuen, S., Billings, D.K., & Ramey, D.R. (1991). *Final report: Longitudinal study of structural immersion strategy, early-exit, and late-exit transitional bilingual education programs for language-minority children.* San Mateo, CA: Aguirre International (Report to the U.S. Department of Education).

Ramos, M., Aguilar, J.V., & Sibayan, B.F. (1967). *The determination and implementation of language policy.* Quezon City, The Philippines: Philippine Center for Language Study: Monograph Series 2.

Reese, L., Garnier, H., Gallimore, R., & Goldenberg, C. (2000). Longitudinal analysis of the antecedents of emergent Spanish literacy and middle-school English reading achievement of Spanish-speaking students. *American Educational Research Journal, 37* (3), 633-662.

Roser, N.L., Hoffman, J.V., & Farest, C. (1990). Language, literature, and at-risk children. *Reading Teacher, 43* (8), 554-559.

Rosier, P. & Holm, W. (1980). *The Rock Point Experience: A longitudinal study of a Navajo school program.* Washington DC: Center for Applied Linguistics.

Ross, S.M., Smith, L.J., & Nunnery, J.A. (1998, April). *The relationship of program implementation quality and student achievement.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Rossell, C.H. (1990). The effectiveness of educational alternatives for limited-English-proficient children. In G. Imhoff (Ed.), *Learning in two languages.* New Brunswick, NJ: Transaction Publishers.

Rossell, C.H. & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the teaching of English, 30* (1), pp. 7-69.

Rossell, C., & Ross, J. (1986). The social science evidence on bilingual education. *Journal of Law and Education, 15*, 385-419.

Rothfarb, S.H., Ariza, M.J. & Urrutia, R. (1987). *Evaluation of the Bilingual Curriculum Content (BCC) Project: A three-year study, final report.* Miami: Office of Educational Accountability, Dade County Public Schools.

Saldate, M., Mishra, S. P., & Medina, M. (1985). Bilingual instruction and academic achievement: A longitudinal study. *Journal of Instructional Psychology, 12* (1), 24-30.

Saunders, W.M. (1998). *Improving literacy achievement for English learners in transitional bilingual programs*. Long Beach, CA: Center for Research on Education, Diversity, and Excellence, University of California.

Saunders, W.M., & Goldenberg, C. (1996). *The effects of a comprehensive language arts transition program on the literacy development of English language learners*. Santa Cruz, CA: Center for Research on Education, Diversity, and Excellence, University of California.

Saunders, W.M., & Goldenberg, C. (1999). *The effects of a comprehensive language arts transition program on the literacy development of English language learners*. Santa Cruz, CA: Center for Research on Education, Diversity, and Excellence, University of California.

Schon, I., Hopkins, K., & Davis, A. (1982). The effects of books in Spanish and free reading time on Hispanic students' reading abilities and attitudes. *NABE: The Journal for the National Association for Bilingual Education, 7* (1), 13-20.

Schon, I, Hopkins, K. D. & Vojir, C. (1984). The effects of Spanish reading emphasis on the English and Spanish reading abilities of Hispanic high school students. *The Bilingual Review, 11* (1), 33-39.

Schon, I., Hopkins, K. D., & Vojir, C. (1985). The effects of special reading time in Spanish on the reading abilities and attitudes of Hispanic junior high school students. *Journal of Psycholinguistic Research,* 14, 57-65.

Secada, W.G., Chavez-Chavez, R., Garcia, E., Munoz, C., Oakes, J., Santiago-Santiago, I., & Slavin, R. (1998). *No more excuses: The final report of the Hispanic dropout project.* Washington, DC: U.S. Department of Education.

Shames, R. (1998*). The effects of a community language learning/comprehension processing strategies model on second language reading comprehension*. Doctoral dissertation, Florida Atlantic University.

Skoczylas, R. V. (1972). *An evaluation of some cognitive and affective aspects of a Spanish bilingual education program.* Doctoral dissertation, University of New Mexico.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15* (9), 5-11.

Slavin, R.E. (1995). *Cooperative learning: Theory, research, and practice* (2nd Ed.). Boston: Allyn & Bacon.

Slavin, R.E. (2003). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31* (7), 15-21.

Slavin, R.E., & Calderón, M. (Eds.) (2001). *Effective programs for Latino students.* Mahwah, NJ: Erlbaum.

Slavin, R.E., Leighton, M., & Yampolsky, R. (1990, April). *Effects of Success for All on the achievement of limited English proficient children.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Slavin, R.E., & Madden, N.A., (1994, April). *Lee Conmigo: Effects of Success for All in bilingual first grades.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Slavin, R.E. & Madden, N.A. (1995). *Effects of Success for All on the achievement of English language learners.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Slavin, R.E., & Madden, N.A. (1999). Effects of bilingual and English as a second language adaptations of Success for All on the reading achievement of students acquiring English. *Journal of Education for Students Placed At Risk, 4* (4), 393-416.

Slavin, R.E., & Madden, N.A. (2000). Research on achievement outcomes of Success for All: A summary and response to critics. *Phi Delta Kappan, 82* (1), 38-40, 59-66.

Slavin, R.E., & Madden, N.A. (2001). *One million children: Success for All*. Thousand Oaks, CA: Corwin.

Slavin, R.E., & Yampolsky, R. (1991). *Effects of Success for All on students with limited English proficiency: A three-year evaluation.* Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Sprick, M.M., Howard, L.M., & Fidanque, A. (1998). *Read Well: Critical foundations in primary reading*. Longmont, CO: Sopris West.

Stallings, J. (1986). Using time effectively: A self-analytic approach. In K. Zumwalt (Ed.), *Improving Teaching* (pp. 15-27). Alexandria, VA: ASCD.

Stebbins, L. B., St. Pierre, R.t G., Proper, E. C., Anderson, R. B., & Cerva, T. (1977). *Education as experimentation: A planned variation model, Vol. IV-A. An evaluation of Follow Through*. Cambridge, MA: Abt Associates.

Stern, C. (1975). *Final report of the Compton Unified School District's Title VII bilingual-bicultural project: September 1969 through June 1975*. Compton City, CA: Compton City Schools.

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

Stuart, M. (1999). Getting ready for reading: Early phoneme awareness and phonics training improves reading and spelling in inner-city second language learners. *British Journal of Educational Psychology, 69* (4), 587-605.

Teschner, R. (1990). Adequate motivation and bilingual education. *Southwest Journal of Instruction, 9,* 1-42.

Thomas, W., & Collier, V. (1997). *School effectiveness for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education.

Thomas, W.P., & Collier, V.P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement.* Santa Cruz, CA: University of California at Santa Cruz, Center for Research on Education, Diversity, and Excellence.

Valladolid, L. A. (1991). *The effects of bilingual education of students' academic achievement as they progress through a bilingual program*. Doctoral dissertation, United States International University.

Van Hook, J., & Fix, M. (2000). A profile of the immigrant student population. In J. Ruiz de Velasco, M. Fix, & B. Chu Clewell (Eds.), *Overlooked and underserved: Immigrant students in U.S. secondary schools*. Washington, DC: Urban Institute.

Vasquez, M. (1990). *A longitudinal study of cohort academic success and bilingual education*. Doctoral dissertation, University of Rochester.

Waxman, H.C., Walker de Felix, J., Martinez, A., Knight, S.L., & Padrón, Y. (1994). Effects of implementing classroom instructional models on English language learners' cognitive and affective outcomes. *Bilingual Research Journal, 18* (3 & 4), 1-22.

Wasik, B.A. & Slavin, R.E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly, 28,* 178-200.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55* (3), 269-317.

Willig, A. (1987). Examining bilingual education research through meta-analysis and narrative review: A response to Baker. *Review of Educational Research, 57* (3).

Wong-Fillmore, L., & Valadez, C. (1986). Teaching bilingual learners. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd Ed.). New York: Macmillan.

Woodcock, R.W. (1984). *Woodcock Language Proficiency Battery*. Allen, TX: DLM.

Yap, K.O., Enoki, D. Y., & Ishitani, P. (1988, April). *LEP student achievement: Some pertinent variables and policy implications*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

Yeung, A.E., Marsh, H.W., & Suliman, R. (2000). Can two tongues live in harmony? Analysis of the National Education Longitudinal Study of 1988 (NELS88) longitudinal data on the maintenance of home language. *American Educational Research Journal, 37* (4), 1001-1026.

Zirkel, P.A. (1972). *An evaluation of the effectiveness of selected experimental bilingual education programs in Connecticut*. Doctoral dissertation, University of Connecticut.

Appendix I
Disposition of Studies: Language of Instruction

Cited by*	Authors	Remarks
Methodologically Adequate--Elementary Reading		
RB	Alvarez (1975)	
RB	Bacon et al (1982)	
RB	Campeau et al (1975)	
	Carlisle & Beeman (2000)	
RB & W	Cohen (1975)	
	Doebler & Mardis (1980)	
RB & W	Huzar (1973)	
	J. A. Maldonado (1994)	
RB	J.R. Maldonado (1977)	
RB	Morgan (1971)	
RB	Plante (1976)	
RB	Ramirez et al (1991)	
	Saldate et al (1985)	
Methodologically Adequate--Secondary Reading		
RB & W	Covey (1973)	
RB & W	Kaufman (1968)	
Canadian Studies of French Immersion A62		
RB	Barik & Swain (1975)	
RB	Barik et al (1977)	
RB	Bruck et al (1977)+B55	
RB	Day & Shapson (1988)	
RB	Genesee & Lambert (1983)	
RB	Genesse et al (1989)	
RB & W	Lambert & Tucker (1972)	
Students Were Not Learning the Societal Language		
RB	Ramos et al (1967)	Learning English in the Philippines
No Reading Outcomes (Oral Language Only)		
RB & W	Lum (1971)	
RB	Bates (1970)	6 months; no pretest data provided
RB	Elizondo de Weffer (1972)	4 months; no reading outcomes; also preference for English language usage C>E
RB & W	Legarreta (1979)	
RB	Rothfarb et al (1987)	
Pretests Were Given After Treatments Were Under Way		
RB & W	Danoff, Arias & Coles (1977a)	
RB	Melendez (1980)	
RB	Olesini (1971)	
	Rosier & Holm (1980)	
RB	Rossell (1990)	
RB & W	Skoczylas (1972)	Large pretest differences; No separate analysis for Spanish dominant students; more English dominant children in the control group
RB & W	Stern (1975)	
	Thomas & Collier (2002)	Separate studies in Maine & Houston
RB	Valladolid (1991)	
RB	Yap, Enoki, & Ishitani (1988)	
Redundant		
RB	Ariza (1988)	Redundant with Rothfarb (1987)
RB	Barik & Swain (1978)	Redundant with Barik et al (1977)
RB	Cohen et al (1976)	Redundant with Cohen (1975)
RB	Curiel et al (1980)	Redundant with Curiel (1979)
RB & W	Danoff et al (1977b & 1978)	Redundant with Danoff (1977a)
RB	El Paso ISD (1987 & 1990)	Redundant with El Paso ISD (1992)
RB	Genesee, Lambert and Tucker (1977)	Redundant with Genesee et al (1983)
RB	McConnell (1980a)	Redundant with McConnell (1980b)

No Evidence of Initial Equality		
RB	Ames & Bicks (1978)	Large pretest difference; mixed grades and mixed languages
RB	Barclay (1969)	Large pretest differences; 7 months
RB & W	Carsrud & Curtis (1979 & 1980)	Mixed Spanish and English dominant children in the analaysis
RB	Cottrell (1971)	Poorly matched on SES. ANCOVA was used but no pretest data provided
RB	Curiel (1979)	No measure of early academic ability
RB	El Paso ISD (1992)	No measure of early academic ability
RB	Layden (1972)	Large pretest difference in both Spanish and English; 10 weeks
RB	Malherbe (1946)	Lacked information about initial comparability
RB	Matthews (1979)	Lacked information about initial comparability
RB	Powers (1978)	No measure of early academic ability
RB & W	Stebbins et al (1977)	No measure of early academic ability
RB	Vasquez (1990)	No measure of early academic ability
RB&W	Zirkel (1972)	Large pretest differences in Hartford and Bridgeport. No bilingual instruction in New Britain, New London.
No Appropriate Comparison Group		
RB	Prewitt-Diaz (1979)	17 weeks; initial group difference (control group had been in the US for 3 yrs; exp group just arrived from Puerto Rico); large pretest difference
RB	Gersten (1985)	Study of Direct Instruction; No bilingual comparison group
RB	Becker et al (1982)	Not an evaluation of bilingual programs
RB	Burkheimer et al (1989)	Compared actual performance to expected performance, no real control group
RB	de la Garza & Marcella (1985)	Compared Spanish dominant to English dominant; no pretest data
RB	McConnell (1980b)	Compared to a baseline group; No measure of initial comparability
RB	Medina & Escamilla (1992)	Compared Vietnamese TBE to Hispanic Maintenance Bilingual; no reading outcomes
RB	Lampman (1973)	Mixed Spanish and English dominant children in the pretest analysis; only separate analysis for mean gains
RB	Moore & Parr (1978)	Mixed Spanish and English dominant children; also late pretests for grade 1 and 2
	Thomas & Collier (2002)	Separate studies in Oregon and Florida lacked control groups
	Thomas & Collier (1997)	No control groups
Brief Studies		
RB	Balasubramonian et al (1973)	4 months
Unavailable		
RB	Ciriza (1990)	
RB	Educational Operations Concepts (1991a & b)	
RB & W	McSpadden (1979, 1980)	
RB & W	Pena-Hughes & Solis (1980)	Compared paired bilingual and transitional bilingual programs
RB	Teschner (1990)	

* RB=Rossell & Baker, 1996
W=Willig, 1985