14

Language of Instruction

David J. Francis, Nonie Lesaux, and Diane August

For many years, discussion of effective reading programs for English-language learners has revolved around the question of whether and how children's first language should be used in an instructional program. The focus of this chapter is on studies that compare bilingual programs with programs that use only English. The first part of the chapter provides background information, the second presents the methods used for the review, the third and fourth present information on studies with language minority children and heritage language studies, the fifth presents studies of French Immersion, and the remainder provides a summary of the methodological issues and findings.

The following research questions are addressed in this chapter: What impact does language of instruction have on the literacy learning of language-minority students? Is it better to immerse students in English-language instruction, or are there benefits to developing literacy in English as well as in the native language?

BACKGROUND

Program Types

In this section, we define the types of programs reviewed in the chapter and summarize findings from prior syntheses on this topic. When a child enters school with limited proficiency in English, the school faces a serious dilemma. How can the child be expected to learn the skills and content taught at the same time as he or she is learning English? There may be many options, but two fundamental categories of solutions have predominated: programs that provide instruction only in English (English-only) and programs with some native-language instruction (often called bilingual).

English-Only Programs In an English-only setting, English-language learners are expected to learn in English from the beginning, and their native language plays little or no role in daily reading (and other) instruction. Formal or informal support is likely to be given to help them cope in an all-English context. This

support may include help from a bilingual aide who provides occasional translation or explanation; a separate class in English as a second language (ESL) to help build English skills; and/or the use of scaffolded instruction, in which teachers use specific techniques to help English-language learners understand content delivered in English. English-only instruction may involve placing English-language learners immediately in classes containing native English speakers, or it may involve a separate class composed entirely of English-language learners for some period of time until the children are ready to be mainstreamed. These variations may well be important to student outcomes, but their key common feature is the almost exclusive use of English for instruction, supported by English texts.

Many authors have drawn distinctions among different forms of English-only instruction. One term often encountered is *submersion*, most commonly used pejoratively to refer to sink-or-swim strategies in which no special provision is made for the needs of English-language learners. This approach is contrasted with *structured English immersion*, which refers to a well-planned, gradual phase-in of unmodified English instruction relying initially on special techniques to make content delivered in English accessible to English-language learners. In practice, English-only programs are rarely pure types, and in studies of bilingual education they are rarely described.

Bilingual Programs Bilingual education differs fundamentally from English-only programs in that it provides English-language learners instruction in reading and/or other subjects in their native language. In the United States, most bilingual programs involve Spanish for two reasons: the greater likelihood of a critical mass of students who are Spanish speakers, and the greater availability of teachers who are bilingual in Spanish and English, as well as of Spanish materials, compared with other languages.

In transitional bilingual programs, children may be taught to read entirely in Spanish initially and then transitioned to English. Such programs may be early-exit models, with the transition to English being completed sometime within the first 3 years of the elementary grades, or late-exit models, in which children may continue to receive some native-language instruction throughout elementary school to ensure their mastery of reading and content before being transitioned (see Ramírez, Pasta, Yuen, Billings, & Ramey, 1991). In contrast, paired bilingual models teach children to read in both English and their native language from the beginning of their schooling. Willig (1985) calls this model *alternative immersion* because children are alternatively immersed in native-language and English instruction. Within a few years, however, the native-language reading instruction may be discontinued as children develop the skills needed to succeed in English. This approach contrasts with transitional bilingual education models, in which children are first taught to read primarily in their native language and then transitioned gradually to English-only instruction.

Finally, two-way bilingual programs, or dual-language programs, provide reading instruction in the native language (usually Spanish) and English to English-language learners in classrooms where they are integrated with English speakers who also learn both languages (Calderón & Minaya-Rowe, 2003; Howard, Sugarman, & Christian, 2003). Some two-way programs begin reading instruction for English-language learners in the native language and then

add English, often in third grade (with native-language reading continuing along with English after that). Other programs provide reading instruction in both languages from the beginning. The key difference between two-way bilingual and other approaches is that students are expected to develop and maintain literacy in two languages.

Heritage Language Programs A special case of bilingual education is those programs designed to preserve or show respect for the *heritage* language of the participating children. For example, Morgan (1971) studied a program in Louisiana for children whose parents often spoke French at home, but who themselves generally spoke English. Such heritage language programs are included in this chapter if the outcome variable in the study is an English reading measure. It should be noted, however, that these programs address a different language-related issue from that usually addressed by English-only immersion or bilingual education in that the students are already proficient in English.

French Immersion Programs. Finally, although studies of French immersion programs are not directly relevant to the question of the effectiveness of bilingual programs for language-minority students acquiring the societal language, they are important in gaining a broader understanding of the role of the sociocultural context in literacy development. Several Canadian studies of French immersion programs, in which native-English-speaking children are taught entirely or primarily in French in the early elementary years (e.g., Barik & Swain, 1978; Genesee, Sheiner, Tucker, & Lambert, 1976, 1977), have played an important role in debates about bilingual education.

It is important to note the striking differences between the Canadian studies and those conducted with language-minority students acquiring English as a societal language in the United States. These Canadian Anglophone children were learning a useful second language, but not the language for which they would 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. Further, these studies all involved voluntary programs, in which children's parents wanted their children to learn French. Moreover, the children in these studies were generally upper middle class, not economically disadvantaged. Because the French immersion programs were voluntary, children who did not thrive in them could be, and were, routinely returned to English-only instruction. Thus, the children who completed the programs were self-selected, relatively high achievers.

Variability Within Program Type. As is true in most educational research on program evaluation, although the type of program accounts for some variability in practice and student achievement, its level of implementation accounts for far more (e.g., Tivnan & Hemphill, 2005). Thus, it is not surprising that, although program type typically defines broad guidelines for the use of students' native language, the amount of instructional time in which either language is used is generally not accounted for in program evaluation studies. For example, there is evidence that use of the native language is highly variable even within a single program model, depending not only on how language education policy is

interpreted at the district level, but also on teachers' beliefs, interpretations of political contexts, and language skills (Gandara et al., 2000).

In addition, there is great variability within program models in the quality of instruction. In a review of bilingual research, August and Hakuta (1997) conclude that, although research has generally favored bilingual approaches, the nature of the methods used and the populations to which they have been applied have been important. Specifically, the authors conclude that program quality has been the key to positive outcomes for English-language learners. For example, carefully designed structured immersion programs using only English may be effective, but this does not justify sink-or-swim (or submersion) English-only programs. The same holds true for the quality of bilingual programs.

Further, the context in which programs are implemented varies in ways that influence their design and effectiveness. For example, parental and community goals regarding English acquisition and the benefits of bilingualism, parents' socioeconomic status (SES) and educational background, and students' age at arrival and prior academic schooling are likely to influence academic and language acquisition outcomes (for an in-depth discussion of these issues, see Parts I, II, and III, this volume). Policy also has a significant influence on programming. In the United States, for example, some states require that English be the only language used for instruction. However, directives from higher levels may not be embraced by the educators implementing a program, again resulting in differences between the program's design and actual implementation.

Previous Reviews

Views diverge in the United States regarding the value of the use of an English-language learner's first language for instruction. Researchers cite evidence that children's reading proficiency in their native language is a strong predictor of their ultimate English reading performance (August & Hakuta, 1997; Greene, 1997; Willig, 1985), bilingualism does not interfere with academic achievement in either language (Yeung, Marsh, & Suliman, 2000), and children are able to transfer some literacy skills acquired in their native language to the societal language (studies investigating the relationship between first- and second-language literacy are reviewed in chap. 9). Proponents of bilingual education use these findings, together with the belief that teaching children to read in a language in which they are not yet proficient is an additional risk factor for reading difficulties (Snow, Burns, & Griffin, 1998), to argue for initial instruction in the native language while students are acquiring proficiency in a second language. In addition to the hypothesized academic and cognitive benefits of bilingual instruction, advocates of bilingual education 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 of native-language instruction argue that it interferes with or delays English-language development because children have less opportunity for time on task in English (Rossell, 2000). Further, programs that include instruction in the native language have been criticized for relegating children who receive such instruction to a second-class, separate status within the school and, ultimately, within society (Glenn, 2000).

Reflecting this debate, reviews and research on the educational outcomes of students receiving native-language instruction have reached conflicting conclusions. For an early review, Baker and de Kanter (1981) examined more than 300 evaluations of programs designed for second-language learners. To be included in the review, a study had to either employ random assignment of children to treatment conditions or take measures to ensure that children in the comparison groups were equivalent; studies with no control group were rejected. Of the studies initially located, only 28 satisfied the authors' criteria. Baker and de Kanter offer the following conclusion from their review: "The case for the effectiveness of transitional bilingual education is so weak that exclusive reliance on this instruction method is clearly not justified" (p. 1). Rossell and Baker (1996) used the Baker and de Kanter review, as well as the work of Baker and Pelavin (1984), as the basis for their own review, in which they considered studies that evaluated alternative second-language programs. Of the 300 program evaluations read, they found only 72 methodologically acceptable. Their review included only studies of *good quality*, which they defined as having random assignment to programs, or statistical control for pretreatment differences between groups when random assignment was not possible, and applying appropriate statistical tests to examine differences between control and treatment groups. Other criteria included results based on standardized test scores in English and comparison of students in bilingual programs with control groups of similar students. Rossell and Baker conclude that most methodologically adequate studies failed to find transitional bilingual education more effective than programs with English-only instruction: "Thus the research evidence does not support transitional bilingual education as a superior form of instruction for limited English proficient children" (p. 7). It should be noted that the authors of these two studies do not state that English-only instruction is more effective, but merely that bilingual instruction should not be the only approach mandated by law.

Willig (1985) conducted a meta-analysis of the studies reviewed by Baker and de Kanter (1981), making several changes with regard to inclusion criteria. First, she eliminated five studies conducted outside the United States (three in Canada, one in the Philippines, and one in South Africa) because of significant differences in the students, programs, and contexts in those studies. She also excluded one study in which instruction took place outside the classroom. Finally, she excluded one review because it was not a primary study. Her overall conclusion is quite different from that of Baker and de Kanter: "positive effects for bilingual programs...for all major academic areas" (p. 297). However, it should be noted that Willig was asking a fundamentally different question from that explored by Baker and de Kanter. The latter authors addressed whether bilingual education should be mandated, whereas Willig considered a more modest question: whether bilingual education works. As she notes, she conducted a series of comparisons. One set of comparisons examined how bilingual programs with and without ESL instruction compared with submersion programs or programs in which English-language learners are placed in all-English classrooms with no special instructional support. A second set of comparisons examined bilingual programs that included ESL support with immersion programs that also included ESL support. For both sets of comparisons,

Willig concludes that bilingual education works better than the English-only programs with which it was compared.

Greene (1997) performed a meta-analysis of the set of studies cited by Rossell and Baker (1996), but the analysis included only 11 of those 72 studies. In addition to the criteria used by Rossell and Baker, Greene looked at studies that measured the effects of bilingual programs after at least one academic year. If students were not assigned to treatment and control groups randomly, adequate statistical control for this nonrandom assignment was defined as requiring controls for individual previous test scores, as well as at least some of the individual demographic factors known to influence those scores, such as family income and parental education. In all, Greene rejected studies cited by Rossell and Baker because they were duplicative of other studies in the review (15), could not be located (5), were not evaluations of bilingual programs (3), did not have appropriate control groups (14),[1] measured bilingual education after a short period of time (2), and inadequately controlled for differences between students assigned to bilingual and English-only programs (25). Among the studies that met the author's standard of methodological adequacy, including all those using random assignment to conditions, Greene found that the evidence favored programs that made use of native-language instruction (average effect size 0.21).

Finally, Slavin and Cheung (2004) conducted a best-evidence synthesis, an approach that uses a systematic literature search, quantification of outcomes and effect sizes, and extensive discussion of individual studies that meet inclusion criteria. Seventeen studies met their inclusion standards. They found that, "among 13 studies focusing on elementary reading for Spanish-dominant students, 9 favored bilingual approaches on English reading measures, and 4 found no differences, for a median effect size of +0.45. Weighted by sample size, an effect size of +0.33 was computed, which is significantly different from zero ($p < .05$)" (p. 2).

Differences in study outcomes can be attributed, in part, to differences in the questions asked, the criteria for including studies, and the methods used to synthesize findings. With regard to the research questions asked, for example, the nature of the samples differed depending on the question (e.g., Willig eliminated studies conducted outside the United States, whereas Baker and de Kanter did not). Standards for methodological rigor also differed across the reviews (e.g., Greene eliminated 61 studies that had been included by Rossell and Baker). Only two of the authors (Greene, 1997; Willig, 1985) used meta-analytic techniques and therefore took into account the program effects found in each study, even if they were not statistically significant. As Greene points out, "simply counting positive and negative effect sizes is less precise than a meta-analysis because it does not consider the magnitude or confidence level of effects" (p. 11). In fact, simple vote-counting procedures are known to be conservatively biased, and the magnitude of the bias increases as the number of studies increases (Lipsey & Wilson, 2001).

Of note is that differences in study conclusions were not large. Many reviews that have been labeled as anti-bilingual education found not that use of the native

[1] Greene asserts that, in most of these cases, children in the control group also received some native-language instruction.

language was worse than English-only instruction, but merely that there were no overall differences. The two reviews favorable to bilingual instruction found differences in favor of native-language instruction, but the effect sizes were small to moderate. Of interest is Willig's (1985) conclusion that the better the technical quality of the study was (e.g., if a study used random assignment as opposed to creating post hoc comparison groups), the larger were the effects. This observation raises an interesting possibility: The effectiveness debate may really be carried on at the relatively superficial level of a study's technical quality.

Although the authors of reviews may disagree on the effectiveness of bilingual education, they do not disagree about the overall quality of the available studies. All had to eliminate large numbers of studies from their reviews A flaw in many studies is the failure to equate experimental and control groups on important variables. In some instances, for example, students in the control groups were those who had exited from bilingual programs (Stern, 1975); in other instances, students in the control groups were those who had never needed bilingual services. Willig (1985) found that in the latter cases, the mean effect sizes for the bilingual groups were among the lowest in her study and favored the English-only groups. When the comparison children did qualify for the program, but were eliminated through the process of random assignment, however, the effect sizes favored the bilingual groups. Language exposure in the neighborhood and school settings can also influence differences between the groups studied. In the studies Willig reviewed, regardless of whether the neighborhood language was English or another language, effect sizes were positive for the bilingual group when both groups (i.e., the bilingual group and the English-only comparison group) had the same neighborhood language. However, when the neighborhood language of the comparison group was English and that of the experimental group was Spanish, little or no differences were found between the two groups. Another study flaw is that, in many cases, the authors do not clearly describe the program characteristics and provide little information about the fidelity or quality of program implementation. Finally, the studies cited in prior reviews have routinely ignored the issue of students within classrooms, a problem which has also been ignored to the prior reviews. We will return to this issue in a later section.

METHODS

Our review includes the methodologically adequate studies that have been cited in previous reviews (e.g., Greene, 1997; Rossell & Baker, 1996; Slavin & Cheung, 2004; Willig, 1985), as well as other studies located in a search of the literature as described later. It is important to note that the methods applied in this synthesis have some important limitations. First, in requiring measurable outcomes and control groups, we excluded case studies and qualitative studies. Many such descriptions exist and are valuable in suggesting programs or practices that may be effective, as well as describing the context in which programs take place (studies of this nature are reported and discussed in chaps. 12 and 16). However, these descriptions do not indicate what children would have learned had they not experienced a particular program. Thus, they are not relevant to the overarching question of program effectiveness that guided the review and meta-analytic work for this chapter. Second, it is important to note that a number

of the studies reviewed took place many years ago, and that both social and political contexts, as well as bilingual and English-only immersion programs, have changed over time. Thus, we cannot assume that all outcomes described here would apply to bilingual and immersion programs today. For example, methods used to coordinate and sequence the use of the two languages are much better developed now, as are methods for scaffolding English instruction.

In this chapter, we focus primarily on research comparing English-only and bilingual reading programs used with language-minority students, with measures of English reading as the outcomes. For these studies, we employed systematic procedures and inclusion criteria and discuss the studies in narrative form, while also computing, where feasible, the effect sizes for individual studies and performing a meta-analysis (Cooper, 1998; Cooper & Hedges, 1994) to compare findings across these studies. We also provide a narrative review of the French immersion studies because, as mentioned in the introduction to this chapter, they are important in gaining a broader understanding of the role of the sociocultural context in literacy development.

Searches

As part of this review, we systematically searched electronic databases for studies that compared some use of the native language with English-only instruction (see chap. 1). In addition, we attempted to obtain every study included in the reviews conducted by Willig (1985), Rossell and Baker (1996), Greene (1997), and Slavin and Cheung (2004).

Appendix 14.A contains a list of all of the studies of reading cited by Willig (1985), Rossell and Baker (1996), Greene (1997), and Slavin and Cheung (2004); it indicates those that were disqualified from this review because they did not meet the panel's criteria for methodological adequacy as outlined next. 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). This does not mean that the overall conclusions of other reviews are incorrect. However, it does mean that the effects of language of instruction on reading achievement were explored by the panel with a somewhat different set of studies from those cited by previous reviews.

Criteria for Inclusion

As described in chapter 1, the studies met the same methodological standards as other experimental and quasi-experimental studies included in the overall report. 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. In some instances, pretreatment covariates were not pretest measures of outcomes, but measures of skills related to the outcomes. That is, it was not necessary that pretest measures of outcomes were available as covariates in nonrandomized studies. Studies without control groups, such as pre- and postcomparisons and comparisons with expected scores or gains, were excluded. No studies were excluded on the basis of level of pretreatment differences.

To be consistent with other previous reviews of the research that compare programs using bilingual instruction with those using English-only instruction, we allowed for a broader time frame and venue for publication. Thus, studies included in this chapter include technical reports, dissertations, and studies predating 1980. In addition to the general inclusion criteria described in chapter 1, the studies reviewed in this chapter met other standards of relevance to the purposes of this chapter:

- The studies compared children taught reading in bilingual classes and those taught in English-only classes, as defined in the preceding section. Studies of alternative reading programs for English-language learners that held constant the language of instruction are discussed in later chapters of Part IV.
- The subjects were language-minority students in elementary or secondary schools in English-speaking countries. Studies in which samples were not composed predominantly of language-minority students or that did not allow an estimate of performance separately for language-minority students were excluded (e.g., Skoczylas, 1972). Studies of other societal languages would have been included if they were analogous to the situation of English-language learners in the United States or Canada (e.g., Turkish children learning to read in Dutch in the Netherlands), but no such studies were found that met our other inclusion criteria. Studies of children learning a foreign language were not included.
- Studies of instruction in heritage languages were also included if they met our other criteria. One such study was identified (Morgan, 1971).
- The dependent variables included quantitative measures of vocabulary and English reading performance, such as standardized tests and informal reading inventories.
- Studies included at least a 6-month span between the onset of instruction and posttests; in these cases, most treatment durations were of at least 1 year.
- Despite their variation with respect to sample and context, Canadian studies of French immersion have been widely discussed and are therefore reviewed in a separate section of this chapter. They are not included in the meta-analysis. As a group, these studies are of high methodological quality and constitute effective program evaluations.

Methods of Rating Studies for Inclusion

Once studies had been selected because they were relevant, two individuals independently reviewed them against our consistent set of standards. The coding rubric for the studies can be found in Appendix 14.B. There were two circumstances in which additional reviewers examined a study: when the primary reviewers disagreed on whether an article should be included, and when the consensus opinion of the reviewers regarding inclusion or exclusion differed from the way an article had been handled in a previously published review (Greene, 1997; Rossell & Baker, 1996; Slavin & Cheung, 2004; Willig, 1985). The final disposition of such studies was determined by consensus of the coders and two

methodological experts.[2] Following these procedures, we arrived at a final set of 20 studies that diverged somewhat from those of previous reviews (see Appendix 14.C). Although many studies that appear in the present chapter also appeared in the four prior reviews, some of the studies in those reviews failed to meet our inclusion criteria (see Appendix 14.A). In addition, some studies judged to meet our criteria had been excluded from one or more of the prior reviews.

Study Characteristics

Twenty studies in our database focused on evaluating the impact of language of instruction on literacy acquisition. In addition to the majority of studies focused on the acquisition of literacy by language-minority students (n = 16), this chapter incorporates findings from one heritage language study and three Canadian French immersion studies. In each of the following sections, studies are organized according to grade level (elementary or secondary). Of the studies focused on language-minority students acquiring the societal language, 14 studies investigated program effectiveness with students in the elementary years and 2 with students in the secondary years. Of all studies, 5 used random assignment to the instructional conditions, and 15 used a matching procedure to compare students receiving some native-language instruction with those receiving English-only instruction.

Finally, in light of our discussions in Parts I, II, and III of the various factors other than language of instruction that influence the development of literacy skills, we provide, to the extent possible for each study, sample characteristics (e.g., age, socioeconomic status [SES], length of exposure to the native and target languages); a description of the program type(s); and, if available, the method used for enrollment in the program(s). For studies that compared program types, but did not employ random assignment to instructional conditions, we describe the matching procedures.

Studies Conducted With Language-Minority Students. Fourteen studies included in our review compared language-minority students in the elementary grades who were taught to read with bilingual or English-only instruction (Alvarez, 1975; Campeau et al., 1975; Cohen, Fathman, & Merino, 1976; Danoff, Coles, McLaughlin, & Reynolds, 1978; De la Garza & Medina, 1985; Doebler & Mardis, 1980–1981; Huzar, 1973; Lampman, 1973; J. A. Maldonado, 1994; J. R. Maldonado, 1977; Plante, 1976; Ramírez et al., 1991; Saldate, Mishra, & Medina, 1985; Valladolid, 1991). These studies were characterized methodologically by random assignment to one of the instructional conditions or by a procedure whereby students were matched on pretest variables, such as reading and oral proficiency, or on pre-reading skills. Two studies (Covey, 1973; Kaufman, 1968) in our review compared language-minority students in the secondary grades who were taught to read with bilingual or English-only immersion approaches. Both studies employed random assignment to one of the instructional conditions.

[2] David Francis, University of Houston; and Tim Shanahan, University of Illinois at Chicago.

Heritage Language Studies. One study we reviewed (Morgan, 1971) examined the effectiveness of a program in which language-minority children received instruction in their heritage language. In this case, the heritage language was French.

French Immersion Studies. Three studies in our review (Barik & Swain, 1975, 1978; Barik, Swain, & Nwanunobi, 1977) evaluated French immersion programs for English-speaking children in Canada. However, because they compared French immersion for English-speaking students with monolingual English instruction or brief classes in French as a second language, these were not evaluations of bilingual education per se.

Computation of Effect Sizes and Synthesis of Findings

When possible, we computed effect size estimates for each study by using the pooled within-group standard deviation and either unadjusted or adjusted posttest treatment and control means, or both when both adjusted and unadjusted means were available. In principle, an effect size is the experimental mean minus the control mean, divided by the standard deviation. When this information was lacking, however, we estimated effect sizes by using information provided by the studies and appropriate conversion formulas provided by Shadish, Robinson, and Lu (1999) and Lipsey and Wilson (2001).[3]

The decision to examine adjusted and/or unadjusted means in these studies merits some discussion. The meta-analysis literature lacks a strong consensus on the choice of posttreatment means for the computation of effect sizes in quasi-experimental studies. The challenge, it seems, is deriving effect size estimates that will compare favorably across the collection of studies—that is, that will allow comparison of apples to apples and oranges to oranges. Because the literature on bilingual education is anything but consistent with respect to the reporting of means and standard deviations and the use of pretreatment covariates, there is no single approach that would have allowed us to estimate effect sizes in the same way for all studies. Often studies reported unadjusted means, standard deviations, and test statistics for adjusted means without providing other information necessary to compute an effect size on the adjusted means. For studies comparing groups on adjusted means, adjustments were not always based on the same covariates across studies, and rarely was the information provided to properly estimate the effect size on the adjusted means. Thus, we settled on the approach of computing effect sizes on unadjusted and adjusted means, when possible, and looking for possible factors that explained variability in the effect sizes. Two studies only reported adjusted means, and one of those studies was a randomized trial; in all, the adjusted means constituted 14% of the reported effect sizes (see Appendix 14.E). Thus, we did not feel there was sufficient variation in the type of mean reported to provide a meaningful test of the moderator variable, and we analyzed the unadjusted means with the exception of the two studies that only provided adjusted means (Alvaree, 1975; Kaufman, 1968).

This chapter incorporates only the effect sizes for those studies in which sufficient information is provided to calculate an effect size.

In some instances, we made assumptions to be able to estimate the effect size when information was lacking. For example, we may have had to assume that the pre- and posttest standard deviations were equivalent because the pretest standard deviation was reported, but not the posttest standard deviation, or we may have had to assume that the treatment and control standard deviations were the same when only one of the two was reported. Finally, in the case of two studies (the Alice, Texas, and Houston, Texas, evaluation studies reported in Campeau et al., 1975; Cohen et al., 1976), we estimated the standard deviations from other studies that had used the same outcome measure at the same grades. More specifically, Campeau et al. reported means and significance tests for gain scores, but did not provide the additional information necessary to derive the posttest standard deviation to be used in the denominator of the effect size. Rather than use the gain score standard deviation, which would be expected to underestimate the posttest standard deviation, we estimated the standard deviation from other studies that used the same outcome measure at the same grades by computing the square root of the average of the pooled within-group variances reported in those studies.

Finally, it must be pointed out that none of the studies reviewed in this section addressed the issue of nonindependence of students who are nested inside instructional units. That is, students who receive their instruction in the same classroom/school/district are not independent, and this lack of independence must be taken into account when computing significance tests. From a practical standpoint, the failure to address this nonindependence in individual studies means that standard errors for individual studies are likely to underestimate the true standard errors, and thus confidence intervals around effect sizes for individual studies should be assumed to be too small. Although the extent of underestimation of standard errors will vary across studies to an unknown degree, we have opted not to judge the statistical significance of individual studies because of their failure to adequately address this issue of nonindependence in their analyses and reported statistics. Because the effect size standard errors are used to weight. The effect sizes in the meta-analsis, this issue also impacts the meta-analysis in an unknown way. Consequently we also examine the efect sizes using a procedure that ignores the standard errors of the individual effect size estimates. Additional details regarding study methodology and effect size computations are provided in Appendix 14.D. We performed a meta-analysis on those studies for which effect sizes could be computed; Appendix 14.E presents the effect sizes. As noted earlier, we also described all the studies included in this chapter as part of a qualitative review. For studies not included in the meta-analysis, we report study outcomes in the narrative review.

STUDIES WITH LANGUAGE-MINORITY STUDENTS

Studies With Elementary School Learners

Three of the studies conducted with elementary school learners used random assignment (Huzar, 1973; Maldonado, 1994; Plante, 1976); the remainder employed a design involving the comparison of a group of language-minority students receiving instruction in their native language with a group of language-minority students receiving no structured support in their native language (Alvarez, 1975; Campeau et al., 1975 [five studies]); Cohen et al., 1976; Danoff

et al., 1978; Dela Garza & Medina, 1985; Doebler & Mardis, 1980–1981; Lampman, 1973; Maldonado, 1977; Ramírez et al., 1991; Saldate et al., 1985; Valladolid, 1991.) In these studies, the groups were generally matched by using either pretest scores on measures of reading and oral proficiency or pre-reading skills.

Studies Using Random Assignment. Plante (1976) conducted a study with Spanish-dominant Puerto Rican children who were attending a New Haven, Connecticut, elementary school. The sample included a group of children who received bilingual education in first and second grades (n = 15) and a group of children who received such education in the second and third grades (n = 16). The control group comprised second and third graders who had received no support or instruction in their native language (n = 10 second graders, n = 12 third graders). The school is described as serving a large percentage of children from low-income families. In this study, children were randomly assigned to the experimental group (a paired bilingual model) or a control group in which no native-language support was offered for Spanish-speaking children. Prior to this 2-year study, there was no native-language support for children in New Haven.

In the paired bilingual experimental condition, the native language of one teacher was Spanish and of another was English. The children in this condition were taught all their basic skills (reading, writing, math, science, social studies) in Spanish by the native-Spanish-speaking teacher while receiving instruction in English (an aural–oral approach) from the native-English-speaking teacher. The latter instruction was designed to transition the children to English-only instruction. When an individual child's oral English vocabulary was sufficiently developed, the teacher initiated reading and writing of English.

In addition to random assignment to conditions, equivalence between the experimental and control groups prior to the onset of the bilingual instruction was established on the basis of measures of oral vocabulary in Spanish and English. Plante also conducted attrition analyses, through which it was determined that attrition (n = 14 from the experimental group, n = 5 controls) did not change the arithmetic means on pretests of reading and language, and that if a chance advantage did exist it would favor the control group.

A similar study (Huzar, 1973) was conducted in Perth Amboy, New Jersey, in a school district where children had been randomly assigned to bilingual or English-only instructional conditions. The children in the experimental condition were second-grade (n = 41) and third-grade (n = 43) Spanish-dominant Puerto Rican children who had received bilingual instruction since first grade. These two groups were compared with control groups of second (n = 40) and third (n = 36) graders with similar backgrounds receiving English-only instruction in the same school. Despite random assignment to bilingual education, Huzar also obtained school district data and determined group equivalence on measures of IQ, SES, and initial achievement on a standardized measure of kindergarten readiness. As in the Plante (1976) study, the students in the experimental group were exposed to a paired bilingual instructional model, and thus had two teachers. One teacher taught reading in Spanish for 45 minutes daily, and the other teacher taught reading in English for the same amount of time. Students in the English-only classes received 45 minutes of English reading instruction daily. The author reports that all teaching procedures and quality of instruction were the same for both groups.

Studies With Random Assignment: Elementary Children With Learning Disabilities. Maldonado (1994) carried out a small randomized study involving language-minority students who were in special education classes in Houston, Texas. Twenty second- and third-grade Spanish speakers with learning disabilities were randomly assigned to one of two groups: a bilingual group that was taught mainly in Spanish for a year with a 45-minute ESL period, and a control group that received traditional special education in English. During the second year, half of the instruction in the bilingual program was in English and half in Spanish. In the third year, instruction was primarily in English. The students in the two groups had similar characteristics, including age, education, experience, learning disability, language proficiency, and SES. Children's achievement was assessed at pre- and posttest with a standardized measure of language and reading achievement (California Test of Basic Skills [CTBS]). Information reported in the article is inconsistent and leads to widely varying estimates of the effect of billingual instrutic. These problems are discussed in Appendix 14.D.

Studies Using Matching De la Garza and Medina (1985) conducted a study comparing the reading achievement of a group of Spanish-speaking Mexican children in a transitional bilingual education program (n = 24) and a group of Spanish-speaking children receiving English-only instruction (n = 118). The study was conducted in Tucson, Arizona, with children of low SES, as evidenced by the majority of the sample's qualification for a free or reduced-price lunch program. In the transitional bilingual program, instruction was in Spanish 75% of the time in first grade, 70% of the time in second grade, and 50% of the time in third grade. Most children in the bilingual program transitioned into English reading in third grade. No details are provided on the number of classrooms per grade or on whether the bilingual and English-only classrooms were in the same or different schools.

The children were followed from first through third grades and assessed on measures of reading vocabulary and comprehension at the end of each year. The students in the sample are those who had data available for 3 years. Students in the bilingual program were required to have data available in both English and Spanish; students in the control sample were required to have data available in English. There are several methodological issues related to the study. First, there was no attempt to determine whether those in the sample at any one grade were comparable to those missing at that grade; that is, there was no assessment of bias due to possible differential attrition. Second, although students in the two groups were similar in ethnicity, grade level, duration of program participation, and SES, in fact 94% of the control group was rated as English dominant in first grade. Thus, although both groups consisted of language-minority students, the control and bilingual education groups were not equivalent in English-language proficiency.[4]

Alvarez (1975) conducted a study with 147 Mexican American children of low SES attending two schools in Austin, Texas. Seven classrooms and teachers were included in the total sample. The sample at each school comprised a

[4]This may be a problem with other studies of young learners that is undetected. Although students may be matched on a number of variables at pretest, language dominance is important and is generally not reported.

group of children receiving instruction in Spanish and a group receiving all instruction in English; the children were followed from first to second grade. At the time of the study, bilingual education was optional, and its aim in the primary grades was to emphasize instruction in the child's native language through oral, reading, and writing activities. Simultaneously, oral English-language development was a focus, with the goal of developing sufficient proficiency so that English reading and writing instruction would be possible. By second grade, the bilingual classrooms are described as a balanced combination of Spanish and English, with reading instruction in both languages. The bilingual program paired a native-Spanish-speaking and a native-English-speaking teacher, who shared two classrooms of children.

One of the most widely cited studies of bilingual education is a longitudinal study by Ramírez et al. (1991) that compared Spanish-dominant students in English immersion schools with students receiving two forms of bilingual education: early exit (transition to English-only instruction in Grades 2–4) and late exit (transition to English-only instruction in Grades 5–6). According to a review of the Ramírez et al. (1991) study carried out by the National Research Council (NRC; Meyer & Fienberg, 1992): "All three programs were intended for students who speak Spanish, but have limited ability to speak English. All three programs had, as one of their goals, teaching students English" (p. 67). A group or cohort was followed, beginning in kindergarten, for each of the three programs. For immersion and early-exit programs, an additional cohort was followed beginning in first grade; for late-exit programs, a cohort was followed beginning in third grade.[5]

Schools from nine districts were involved overall, with five sites providing English immersion programs and five sites providing early-exit programs. Late-exit programs were not located in the same districts as English immersion or early-exit programs. English immersion and early-exit programs were generally in the same districts and in four instances were in the same schools. Although within-site comparisons provide for a better test of English immersion versus early exit, including sites that do not involve both program types adds 16 schools in English immersion and 12 schools in early exit, many of which (4 English immersion and 7 early exit) are in the same districts as schools with both programs. Thus, we include estimates of English immersion versus early exit both within and between schools.

Meyer and Fienberg (1992) found that the most compelling findings were from the K–1 analyses comparing the four schools that provided both early-exit and English immersion programs. Children in the two programs were well matched on kindergarten pretests, SES, preschool experience, and other factors. These authors did not think that late-exit versus English immersion comparison

[5]Meyer and Fienberg (1992) found three comparisons unacceptable: those using the first-grade cohort because no information is provided about the type of program the students attended prior to first grade; comparisons between early-exit bilingual programs and immersion programs located in different schools because, even after including the background variables in the model, statistically significant school effects were found; and comparisons of the late-exit model with the other two models because the districts in which the late-exit model were implemented did not have the other two kinds of models, making it impossible to compare students in the late-exit programs with those in other programs while controlling for district differences.

was warranted because of differences in sites and school-level heterogeneity that was confounded with programs. However, these factors are operating in other comparisons included in our analysis. Thus, it seems reasonable to examine the English immersion versus late-exit comparisons even when the programs are located in different districts schools. Because Grade 3 is the highest grade available for any of these comparisons, no data are reported beyond that grade.

In a small study conducted in New Mexico, Lampman (1973) examined the academic achievement of 40 Spanish-speaking second graders in bilingual classrooms (n = 20) and mainstream classrooms in which English was the language of instruction (n = 20). The children in the study were matched on age, IQ, home language practices, and demographic variables. At the end of second grade, there were no differences between the two groups using grade equivalent scores. This study was not included in the meta-analysis because the authors reported only mean grade equivalent scores and did not provide sufficient information to compute an effect size estimate.

Saldate et al. (1985) studied 62 children in an Arizona border town who attended English-only or bilingual programs. The participants in the bilingual program were Mexican American children of low SES as indicated by the location of the school they attended. The children in the English-only program were from nearby schools in the same district serving mainly Mexican Americans (60%–90%). Spanish-speaking students in the experimental group were enrolled in a bilingual/bicultural program whose goals included development of Spanish and English literacy, improvement of cognitive functioning, enhanced knowledge of Mexican and American cultures, and development of positive self-concept and motivation for learning. In first grade, the children were individually matched on a standardized measure of vocabulary and placed into pairs of experimental and control subjects. Students were followed into third grade and assessed on English and Spanish reading achievement tests. Given the small sample size of this study, the results should be interpreted cautiously, especially because the number of pairs in the analysis dropped from 31 to 19 between second and third grades, and no attrition analyses are presented. Also, the study is designed as matched pairs, but data are analyzed as independent groups. Nesting of students within classrooms is ignored as is the case with the majority of studies in this chapter.

Valladolid (1991) conducted a study to determine whether bilingual education had an impact on Hispanic language-minority students' academic achievement compared with a group of students receiving English-only instruction. The study included 107 Hispanic students who had been enrolled in a California school district from kindergarten through Grade 5. Fifty-seven of the students had been enrolled in a bilingual program throughout their schooling and 50 in a traditional English-only program. Both experimental and control groups consisted of students with similar language proficiency and background characteristics. Before students were placed in one of the two types of classes, parents and guardians were informed of the bilingual classes, and students of parents who opposed bilingual education were placed in the traditional English-only program. The bilingual program was driven by the goal of developing proficiency in the basic skills of listening, speaking, and writing in the students' native language so that these skills would transfer to the second

language. Second-language vocabulary was introduced in an ESOL program. Once children transitioned into English reading (generally in third grade), they transitioned into English-language arts, math, and other academic programs. Those bilingual children identified as having limited English proficiency who were enrolled in traditional classes received daily structured lessons in English, provided on a pull-out basis by instructional aides; they also did supplementary work with English reading teachers.

Maldonado (1977) conducted a study with Mexican American children enrolled in bilingual and English-only classes in schools in Corpus Christi, Texas. The experimental group comprised children who had been enrolled in a bilingual program for 4 consecutive years—in the first, second, third, and fourth grades. The control group comprised children who had never received bilingual instruction from first through fourth grades. The students were from families of low SES. At fifth grade, all students were in a mainstream setting. First-grade reading scores were used as a control variable.

One large-scale program evaluation study was the Impact Study of the Elementary and Secondary Education Act Title VII Spanish/English Bilingual Program (Danoff et al., 1978), designed to evaluate bilingual education projects funded by the U.S. Office of Education. The study was designed to contrast the performance of students enrolled in Spanish–English bilingual programs receiving federal Title VII funds with comparable students not enrolled in such programs. During the 1975–1976 school year, students in Grades 2 to 6 in each group were pre- and posttested, and a subsample (the Follow-On Sample) of those students in second and third grades was also tested in the fall of the following year. The following procedure was used to select the sample for the study.

> From the total pool of Title VII classrooms in each of the thirty-eight project sites, a stratified random sample was drawn which included at least one classroom for each site from every grade second through sixth; to the extent that participating sites would agree, additional classrooms were randomly chosen so that approximately 40% to 50% of the Title VII classrooms in each participating site were tested. In addition, non-Title VII classrooms were selected in 20 sites which were able to nominate non-Title VII classrooms within or near their district whose students were comparable to Title VII students in terms of ethnicity, socio-economic status, and grade levels. (p. 3)

Only 75% of the students enrolled in the bilingual classrooms were of Hispanic origin.

In all, 5,311 treatment students and 2,460 control students participated in the Impact Study; for the Follow-on Sample, there were 191 Title VII second graders, 63 non-Title VII second graders, 201 Title VII third graders, and 81 non-Title VII third graders. As mentioned earlier, the authors state that the comparison group was selected by matching Title VII program students with mainstream students within or near the district by ethnicity, SES, and grade levels. It should be noted that the students who participated in the study were selected from 11,073 students in second through sixth grades in 150 schools from 38 school districts: 7,364 students from Spanish-speaking backgrounds who were enrolled in Title VII-funded programs and 3,709 students in non-Title VII classrooms. Because of the large scope of this study, drawing from multiple school districts, it is likely

that policies for enrollment in bilingual education programs, as well as the characteristics of the programs, varied from district to district.

Pooled within-group standard deviations for unadjusted posttests were available to compute effect sizes for both the unadjusted and adjusted posttest means. Nesting of students within classrooms, schools, and sites was ignored; therefore, standard errors are underestimated.

The percentages of students in the Title VII sample who had spent their entire schooling in bilingual classrooms were 40%, 35.4%, 26.1%, 17.9%, and 8.7% in Grades 2 to 6, respectively (Danoff et al., 1978). These percentages reflect the number of students who started bilingual instruction in kindergarten (assuming no grade repeats). More than 20% of the sample at each grade had spent 1 year or less in bilingual classrooms, with a high of 31.3% (Grade 2) and a low of 20.7% (Grade 5). At the same time, the authors report that programs generally tended to keep students in the bilingual classrooms once the students could function fully in English (Danoff et al., 1978). This claim seems to be at variance with the percentage of students with consistent experience in bilingual classrooms.

Further, there was differential attrition from fall to spring across the two groups. Attrition was consistently higher in non-Title VII classes, ranging from 40% to 17%, in contrast with 11% to 22% for Title VII classrooms. In all but Grade 4, the differences are relatively substantial. Although the authors state that rates of attrition were not dramatically different, they were (12%, 7%, 3%, 16%, and 18% in Grades 2–6, respectively). Given the large sample sizes (n = 158–1,370) and the large overall attrition in the non-Title VII classrooms, the differences in attrition rates seem to warrant examination for differences between those who remained in the sample and those who did not in the two groups. Appendix 14.C reports that students missing at the follow-up tended to have lower scores at the pretest than students present at both time points. Nevertheless, despite the differential attrition in all but Grade 4, the authors conclude that this effect was not likely to bias the results of the trends reported on growth.

Cohen et al. (1976) conducted a longitudinal study with Mexican American first through fifth graders of low SES. The study included three cohorts, each followed for 3 years (Grades 3–5, Grades 2–4, and Grades 1–3). Although the bilingual program was implemented in only one school, there was extensive variability in its implementation from year to year and from grade to grade. In all grades and in all years, however, teachers and aides used both English and Spanish in math, social studies, and science lessons, even at the initial stages of instruction, so that children were learning in both languages simultaneously.[9] The treatment sample was matched with students in a nearby school in the same community also received. English-only instruction, with approximately half of the comparison group also receiving special attention through ESL or Title I instruction, as well as individual tutorials. Children were tested during each year of the study. The authors note that some of the control students were spending summers in Mexico, where they may have learned to read in Spanish.

[9]According to the authors, "this generally meant that Spanish and English were used interchangeably (word for word, phrase for phrase, sentence for sentence) or one after the other" (pp. 3–4).

In the study, performance trends over time are based on children who remain in the cohort. The data presented in Cohen et al. (1976) raise concerns about the effects of attrition in both the treatment and control groups. The means of the scale scores show an inconsistent trend, going up and down over time with ever-decreasing sample sizes. Unfortunately, there is no analysis of attrition effects, and thus the extent to which the patterns relate to loss of subjects rather than measurement error or to changes in subjects' ability cannot be ascertained. We also note that the overall sample sizes for each group were small, ranging from 14 to 7. Questions of both the magnitude of overall attrition and whether attrition was differential across the groups hinder interpretation of the study's findings. As in most other studies in the review, the analysis does not take into account nesting of students and effects of higher level nesting units (e.g., schools) in the analysis, limiting interpretation of reported significance tests. Finally, like many other studies in this review, Cohen et al. (1976) used the Inter-American Reading Test (IART) at different grade levels. Different levels of the test were used in each year/grade of the study and are footnoted in the text as Levels I, II, and III. The three forms of the test appear not to be equated, indicating that the trend in means over time is due to changes in ability as well as changes in the test, in addition to the effects of attrition noted previously. Also as noted, the published report does not provide sufficient information to estimate effect sizes directly. Rather, we had to use estimates of the standard deviation for the IART at each grade from other studies to estimate the effect sizes for this study.

Doebler and Mardis (1980–1981) compared a bilingual program in Choctaw with English-only instruction among 63 Choctaw second graders in Mississippi. All the subjects were native Choctaw speakers, and none was fluent in English. Exposure to English occurred only in the classroom because children spoke Choctaw at home and on the playground. It should be noted that all students had been taught in Choctaw with ESL instruction in kindergarten and first grade. Seven classrooms participated in the study—four experimental and three control. The decision to participate as a bilingual or control classroom was left to the staff at each school. The bilingual program taught mathematics, reading, and science in the Choctaw language, with supplementary ESL instruction to teach English reading and language arts and reinforce content concepts taught in Choctaw. In the control condition, children were taught solely in English by certified teachers. Controlling for performance on a standardized measure of reading in English administered in the fall, there were no differences between the groups on the same measure in the spring of second grade. The analysis reported with the study did not take into account assignment at the classroom level, but instead treated students as the unit of assignment and, like other studies in the review, did not adjust standard errors for nesting effects. Finally, the study did not report sufficient information to allow estimation of the effect size and thus had to be excluded from the meta-analysis.

Exemplary Bilingual Programs. In the mid-1970s, the American Institutes for Research (AIR) produced a report on bilingual programs around the United States (Campeau et al., 1975). The studies included in that report are of interest, with the caveat that the AIR researchers were looking for exemplary bilingual

programs. They began with 175 candidates and ultimately winnowed this number down to studies of 7 programs. Four studies in the report met our criteria for this review; they are described later. Three studies were excluded for the following reasons: Maine (no control group), Philadelphia (no control group), and Kingsville (no student outcome data reported).

A study in Corpus Christi, Texas, evaluated a bilingual program in three schools. The study was conducted with Mexican American native speakers of Spanish of low SES. The kindergarten program developed both English and Spanish oral language and reading readiness skills in the students, but the emphasis was on Spanish (90% of the instruction). In first grade, about 1 hour was devoted to Spanish reading and language arts and 2 hours to English reading and language arts. During Grades 2 to 4, Spanish and English reading continued to be developed. Bilingual teachers were used exclusively in kindergarten and first grade; in Grades 2 to 4, a paired model was used. The control group consisted of students in three different schools who received all their instruction in English. The students were approximately equal to control students with regard to SES. Equal numbers of students in both groups were native Spanish speakers (74%). In the 1972–1973 cohort, experimental and control classes were matched on both English and Spanish measures. Because results of kindergarten pretests for these first graders are not given, the findings should be interpreted with caution because attrition over 2 years could have rendered the initially equivalent samples unequal. A second kindergarten cohort (1973–1974) receiving bilingual education was also compared with a control group receiving English-only instruction.

Another study included in the Campeau et al. (1975) report was conducted in Houston, Texas. Three cohorts of students in seven bilingual[7] and two English immersion schools were followed from kindergarten through third grade. The authors reported that "control groups were selected based on their similarity to the experimental groups in language, socio-economic level, and academic achievement" (p. 157). Instruction included a block of time devoted to Spanish reading and language arts. During the remainder of the day, instruction was in English for English-dominant and bilingual students. Spanish-dominant students received additional instruction in Spanish after the lessons had been presented in English. The authors note that attrition over the 4 years of the program was significant. For example, just 75 of the 290 kindergarten pupils enrolled in the bilingual program in 1969–1970 remained in the program in third grade.

A third study in Alice, Texas, also included in the Campeau et al. (1975) report, compared children placed in bilingual programs because of "English language problems, parent approval, and sufficient space" (p. 127). One control classroom at each grade level was composed of children whose oral language eligibility test scores matched those of the bilingual sample most closely. The authors report that no control student later entered a bilingual program or vice versa. The 1972–1973 group included 397 students in bilingual programs in Grades K to 3 and 102 control students; the 1973–1974 group included 504

[7]In 1973–1974, there were eight bilingual schools.

treatment students in Grades K to 4 and 136 controls. The bilingual program began with a focus on Spanish literacy skills in kindergarten, with some language arts instruction in English. By January of first grade, all children participated in reading instruction in English, and through to fifth grade, instruction was in equal amounts of Spanish and English. The authors note that some teachers taught 1 week in Spanish and the next week in English; others alternated the two every other day.

Finally, a 1-year study carried out in Santa Fe, New Mexico, also included in the Campeau et al. (1975) report, examined the reading achievement of children in Grades 1 to 4. "The bilingual program added to the regular English program, a Spanish instructional component that complemented and reinforced the instruction in all content areas. Thus, students received a bilingual presentation of all the topics of study in the normal curriculum" (p. 92). In this particular district, parents chose whether to place their children in bilingual or English-only programs. Pretest scores were higher in the bilingual program in first grade, but not in the other three grades. Within each of the grade levels, a comparison was made from fall to spring of the given year.

As noted, the programs studied by Campeau et al. (1975) are not representative of all bilingual programs because the authors focused by design on exemplary programs. A potential confound, moreover, is that we have no information about the English-only programs. If they were of inferior quality, the positive effects found for the bilingual programs may have been due to those programs' excellent instructional methods, rather than the language of instruction. Because several of these studies did have well-matched control groups and met our review criteria, however, they were included in this review.

Studies With Secondary School Learners

Two studies qualifying for our review evaluated programs that introduced Spanish-language instruction to language-minority students in the secondary grades. Both used random assignment to conditions.

Covey (1973) randomly assigned to bilingual or English-only instructional conditions 200 Mexican American ninth graders attending an urban high school in the southwestern United States. The students were selected from a group of 379 students who had initially been identified to participate. "To be included in the study, students had to demonstrate limited ability to speak English, come from a bilingual home, manifest a reading deficiency, and possess a deficiency in English and mathematics" (p. 56). The experimental intervention (i.e., the instructional techniques used with the students) is not described in any detail. The author defines *bilingual education* as "the use of two languages, one of which is Spanish and the other English, as mediums of instruction for the same student population in an organized instructional program, consisting of English, mathematics and reading," and a *regular program* as one in which "one language is used for the medium of instruction for the same student population in a well organized program which encompasses English, mathematics and reading" (p. 14). No further information is provided about the programs. The groups' scores were nearly identical at pretest on the Stanford Diagnostic Reading Test, as expected given assignment at random to

treatment and control. However, the random assignment process is not discussed, making it impossible to evaluate the assignment process independently of the observed pretest mean equivalence.

It should be noted that pretests were not used as covariates in the analysis of posttests, lowering overall power for the tests of treatment effects reported in the study. At the same time, the fact that the study failed to take into account the nesting of students for instruction would lead to underestimation of standard errors, a problem affecting all other studies reported on in this review, and one that would have the opposite effect on power. It should also be pointed out for potential future reviewers that the reported analyses of within-group pre- and postchanges are incorrect as reported, in that they do not take into account nonindependence of observations on the same students over time. This problem also leads to an inflation of the Type I error rate. Although the analyses reported in Covey (1973) are incorrect, the report includes sufficient information to estimate the effect size for use in the meta-analysis. That is, the problems with the reported analyses do not affect the estimate of the standardized posttest mean difference, although the nesting problem will tend to lead to underestimation of the standard error of the effect size. Again, this problem was present in all reviewed studies.

Kaufman (1968) evaluated a program in which low-achieving Spanish-speaking seventh graders were randomly assigned to bilingual or English-only instructional conditions in two New York junior high schools. One school participated in the program for 1 year and one participated for 2 years. As the author notes, "at each school, students in the treatment and control groups received equivalent instruction in English" (p. 523). For 45 minutes a day (3 days a week in School B and 4 days a week in School A), however, students in the treatment group received instruction in standard Spanish, with emphasis on specific reading skills in Spanish, whereas the control group received extra periods of art, music, and health education conducted in English. Some people have criticized this study because students in the bilingual group received additional instruction in Spanish focused on reading, whereas students in the control group received additional instruction in English, but focused on music, art, and health. The criticism assumes that the appropriate control is to provide an equivalent amount of additional time in English literacy instruction. However, if the study were designed in this way, the groups would not have comparable amounts of English literacy instruction. The analyses reported in the study included adjustment of posttest means for covariates other than the pretest, including language-based IQ, nonverbal IQ, age, and capacity. The study did not report unadjusted posttreatment means, however; because of random assignment, these would be expected to equal the adjusted posttreatment means in the long run. Unfortunately, the study reported outcomes in terms of grade equivalent scores and failed to report information on clustering of students in classrooms. It should also be noted that the average grade equivalent scores for students were roughly 3 to 4 years below the current grade-level placement in both groups.

HERITAGE LANGUAGE STUDIES

As noted, our review also included one study that examined the effectiveness of programs in which children who are proficient in the societal language (English

in the United States) receive instruction in their heritage language. Typically, these are children whose parents also speak the societal language in the home, but would like their children to be fluent in their heritage language as well.

Morgan (1971) carried out a study with almost 200 children of French-speaking parents in rural Louisiana. Fifty-four first-grade classes made up the population. Classes were either bilingual (16) or English (38) depending on the teacher's competence in French. The bilingual group participating in the study included all the students in bilingual classes whose parents scored above the median on a questionnaire crafted to assess level of proficiency in French (93 students); the monolingual group consisted of 100 students in English-only classes randomly selected from a pool of 199 students whose parents spoke above the median level of French. The first graders were followed for 1 year. In the bilingual classes, children were taught in both French and English. The bilingual program was designed to teach French through the oral–aural approach, and French cultural appreciation was developed through songs, plays, and real objects. Formal, structured French-language instruction was conducted for a 30-minute period each day as part of the 2—hour period of language arts instruction; the remaining 90 minutes were devoted to English language arts. All other basic instruction was in English, but casual conversation in French was allowed and encouraged. Children in the monolingual group received all of their instruction in English, and French conversation was not encouraged. These children also received 120 minutes of language arts instruction in English. At the beginning of first grade, the two groups were virtually identical on English tests of mental abilities and readiness. At the end of first grade, students were compared on four English reading measures.

SUMMARY

To evaluate the impact of bilingual education as compared with English-only instruction, we analyzed the estimated effect sizes from the 15 studies by using the *Comprehensive Meta-Analysis (CMA) Version 2* software (Borenstein, 2005). Appendix 14.E provides a table with results for each study, sample, outcome, and grade that went into the meta-analysis. For all studies, positive effect sizes indicate a difference favoring bilingual education, whereas negative effects indicate a difference favoring English-only instruction. In estimating the average effect size, we first corrected the reported effect sizes for small-sample bias; that is, we converted the effect sizes to Hedges' g^U through the CMA software (Hedges, 1981). Each effect size was also weighted by the inverse of its variance, which is a function of both sample size in the treatment and control groups and the effect size. In averaging across effect sizes, we treated each study sample as the unit of analysis. Thus, the 15 studies yielded 71 effect sizes across 26 samples. For the sake of computing average effect sizes, we averaged across different reading outcomes and grades within the same study sample to derive a weighted average for that study sample. These weighted average effect sizes for each study sample appear in Table 14.1,[8] along with their estimated standard errors and 95%

[8]Tables appear throughout chapter.

TABLE 14-1.
Effect size statistics for individual studies

			Statistics for Each Study						
RCT	*Study Name*	*Subgroup Within Study*	*Hedges' g^u*	*Standard Error*	*Variance*	*Lower Limit*	*Upper Limit*	*Z Value*	*p Value*
Yes	Maldonado, 1994	Sample 1	2.1212	0.5440	0.2959	1.0550	3.1874	3.8992	.0001
	Saldate et al., 1985	Sample 1	-0.2829	0.2521	0.0636	-0.7770	0.2112	-1.1223	.2617
	de la Garza, 1985	Sample 1	0.1910	0.2194	0.0482	-0.2391	0.6211	0.8703	.3841
	Ramírez et al., 1991	Sample 1	0.1774	0.1484	0.0220	-0.1135	0.4684	1.1953	.2320
	Ramírez et al., 1991	Sample 2	0.0947	0.0954	0.0091	-0.0923	0.2817	0.9930	.3207
	Ramírez et al., 1991	Sample 3	0.0796	0.1049	0.0110	-0.1259	0.2852	0.7591	.4478
	Valladolid, 1991	Sample 1	-0.6052	0.1968	0.0387	-0.9909	-0.2196	-3.0758	.0021
	Alvarez, 1975	Sample 1	-0.1863	0.2390	0.0571	-0.6548	0.2822	-0.7795	.4357
	Alvarez, 1975	Sample 2	-0.2541	0.2389	0.0571	-0.7224	0.2142	-1.0634	.2876
	Campeau et al.,1975	Sample 2	1.8279	0.2426	0.0589	1.3523	2.3034	7.5340	.0000
	Campeau et al., 1975	Sample 3	1.3929	0.2628	0.0691	0.8778	1.9080	5.2999	.0000
	Campeau et al., 1975	Sample 5	2.6311	0.2230	0.0497	2.1941	3.0681	11.8001	.0000
	Campeau et al., 1975	Sample 6	0.2420	0.1357	0.0184	-0.0239	0.5080	1.7837	.0745
	Campeau et al., 1975	Sample 7	0.8540	0.1585	0.0251	0.5434	1.1646	5.3889	.0000
	Campeau et al., 1975	Sample 8	0.4553	0.1716	0.0294	0.1191	0.7916	2.6540	.0080
	Cohen et al., 1976	Sample 1	-0.1741	0.3904	0.1524	-0.9392	0.5911	-0.4459	.6557
	Cohen et al., 1976	Sample 2	-1.1518	0.4591	0.2108	-2.0516	-0.2519	-2.5087	.0121
	Cohen et al., 1976	Sample 3	-1.5981	0.5539	0.3068	-2.6838	-0.5125	-2.8851	.0039
	Danoff et al., 1978	Sample 1	-0.2621	0.0690	0.0048	-0.3974	-0.1269	-3.7992	.0001
Yes	Huzar, 1973	Sample 1	0.0136	0.2201	0.0485	-0.4178	0.4451	0.0619	.9506
Yes	Kaufman, 1968	Sample 1	0.0477	0.2355	0.0555	-0.4139	0.5092	0.2025	.8396
Yes	Kaufman, 1968	Sample 2	0.4696	0.2989	0.0893	-0.1161	1.0554	1.5714	.1161

(Continued)

TABLE 14-1.
(Continued)

			Statistics for Each Study						
RCT	*Study Name*	*Subgroup Within Study*	*Hedges' g*	*Standard Error*	*Variance*	*Lower Limit*	*Upper Limit*	*Z Value*	*p Value*
	Maldonado, 1977	Sample 1	0.3580	0.1845	0.0340	- 0.0036	0.7195	1.9404	.0523
Yes	Plante, 1976	Sample 1	0.7750	0.4097	0.1679	-0.0281	1.5780	1.8915	.0586
Yes	Covey, 1973	Sample 1	0.6583	0.1555	0.0242	0.3534	0.9631	4.2323	.0000
	Morgan, 1971	Sample 1	0.2541	0.1441	0.0208	-0.0283	0.5365	1.7635	.0778

Note: STANDARD errors do not take into account potential effects of clustering within studies. Confidence intervals, *z*-values, and *p*-values should be interpreted with caution.

confidence intervals. These weighted averages were then averaged to estimate the mean effect size and its standard error under each of two models: a fixed effects model and a random effects model. The weighted average across all study samples appears in Table 14.2, along with an estimate of the standard error, the lower and upper limits of a 95% confidence interval, and a test that the mean effect size equals zero. In addition to computing the average effect size across all studies, we also computed the mean separately for the studies that used randomization. This estimate appears in Table 14.2 as well. Finally, because Maldonado (1994) produced a somewhat larger effect size than the remaining randomized controlled trials (RCTs), and because information reported in Maldonado (1994) was internally inconsistent indicating possible errors in our estimate of the effect size, we also computed the mean effect size separately for the RCTs without Maldonado to assess the overall impact of this one large effect size on the mean estimate and conclusion for the RCTs. These estimates appear in the final two rows of Table 14.2.

Scanning Table 14.1 reveals a range of effect sizes from negative to positive, with at least some statistically significant positive and negative effect sizes (i.e., effect sizes in either direction that are statistically different from 0). Overall, 16 of the 26 estimated effect sizes are positive, 8 are negative, and 2 are effectively 0 (i.e., between 0 and .05). At the same time, only 7 of the 16 positive effect sizes have confidence intervals that exclude 0, and only 4 of the 8 negative effect sizes exclude 0. These observations suggest that the effect sizes vary somewhat across the studies in this review, and, in fact, a test for heterogeneity corroborates that conclusion ($Q = 323.7$, $df = 25$, $p < .0001$). Although the weighted average of the effect sizes is significantly different from 0 (mean = .18, $SE = .033$, $p < .0001$ under the fixed effects model; mean = .33, $SE = 0.127$, $p = .011$ under the random effects model; Table 14.2), the test for heterogeneity indicates that the average effect size may not describe very well the collection of effect sizes. Thus, we separately examined those five studies that used random assignment of students to condition.

Separate examination of the five studies that involved randomization (6 samples and 12 individual effect sizes) produced a somewhat larger weighted average effect size that was also statistically different from 0 under both the fixed and random effects models (mean = 0.45, $SE = .11$, $p < .0001$ under the fixed effects model; mean = .54, $SE = 0.21$, $p = .012$ under the random effects model). In addition, the test for heterogeneity again showed that the effect sizes were not consistent across the collection of studies ($Q = 18.7$, $df = 5$, $p = .002$), although in this case four of six effect sizes are positive and two fall between 0 and 0.05. That is, all effect sizes are in the same direction, but they vary somewhat in magnitude. Although these findings suggest a moderate effect of bilingual education, examination of the effect sizes included in this subset analysis indicates one large effect size of 2.12 ($SE = .54$) associated with Maldonado (1994) that we know to be problematic. Results reported in that study are internally inconsistent, in that different results reported for the same outcome and sample give different effect sizes, as described in Appendix 14.D. The effect size used in the analysis is based on the reported means and standard deviations in the paper, but assuming the reported standard deviations were actually standard errors. That is, we multiplied the reported standard deviations by the square root of the sample size. An effect size computed on the reported standard deviations would have been slightly over 7.0 in magnitude, a highly unrealistic result and one not at all consistent with other information

TABLE 14.2
Statistics for Average Effect Sizes

		Statistics for Average Effect Size						
Model	*Studies Include*	*Hedges' g^u*	*Standard Error*	*Lower Variance*	*Upper Limit*	*Limit*	*Z Value*	*p Value*
Fixed	All studies	0.1835	0.0329	0.0011	0.1191	0.2479	5.5838	.0000
Random	All studies	0.3251	0.1271	0.0162	0.0760	0.5743	2.5575	.0105
Fixed	RCTs	0.4515	0.0997	0.0099	0.2560	0.6470	4.5273	.0000
Random	RCTs	0.5380	0.2140	0.0458	0.1185	0.9574	2.5136	.0119
Fixed	RCTs except Maldonado, 1994	0.3934	0.1014	0.0103	0.1946	0.5923	3.8782	.0001
Random	RCTs except Maldonado, 1994	0.3650	0.1638	0.0268	0.0440	0.6859	2.2287	.0258

reported in the paper. Specifically, Maldonado (1994) also reports the obtained *t*-test result of the difference between the means of the treatment and control groups, and this result could not have resulted from the means and standard deviations reported in the text because it is substantially too small. In addition, it appears that the pre- and posttest means for the control group have been reversed. The effect size based on the reported *t* statistic in Maldonado (1994) is still large ($d = 1.72$, compared with $d = 2.25$ prior to correction to Hedges' $g^u = 2.12$). Insofar as it is impossible to determine which reported numbers are in error, although it appears that the standard deviations are too small, we reanalyzed the RCTs after eliminating Maldonado (1994) from the collection of studies to assess the magnitude of the treatment effect in the remaining four randomized trials. Here again the weighted average of the treatment effects indicates a statistically significant, moderately sized, average treatment effect regardless of which statistical model is assumed for the distribution of effect sizes (mean = .39, $SE = .10$, $p < .0001$ under the fixed effects model; mean = .365, $SE = 0.16$, $p = .026$ under the random effects model). In addition, there is some remaining evidence of heterogeneity in the effect sizes ($Q = 8.964$, $df = 4$, $p = .062$). Although it would certainly be possible to take an alternative approach to dealing with the inconsistencies in the reported data for Maldonado (1994), eliminating the study is the most conservative. Eliminating Maldonado from the collection of all 26 effect sizes has a minimal effect on the fixed effect estimate of the average effect size, which drops from .180 to .176, and a slightly larger, but still relatively negligible effect on the random effects estimate, which drops from .33 to .28. Both effects remain statistically significant. Finally, because each of these analyses involves weighting by the standard error of the effect sizes, which we know to be in error because of the likely effects of clustering in individual studies, we also computed one sample *t*-statistics for each of the three collections of studies. These tests completely ignore the standard errors of the effect sizes and rely only on the collection of effect sizes across the studies. The computed *t*-statistics were as follows: All studies $t(25) = 1.70$; RCTs, $t(5) = 1.97$: RCT without Maldonado, $t(4) = 2.26$.

In summary, it seems reasonably safe to conclude that bilingual education has a positive effect on English reading outcomes that are small to moderate in size. The best evidence supporting this conclusion is that taken from the randomized studies, either with or without Maldonado (1994) included in the collection, or included with some adjustment to the estimated effect size to address the inconsistency across the reported results of the study. It seems equally safe to conclude that many questions regarding how to make bilingual instruction maximally effective for students, and the factors that moderate this effectiveness, remain unanswered by studies reviewed in this chapter. We would have liked to conduct an exhaustive examination of potential moderator variables in the analyses conducted here, but, because of limited resources and time, could not do so. To the extent that the studies provided relevant information, such an analysis could yield some benefit to understanding the research literature and instructional effectiveness.

The majority of the studies included in our review employed a matched design, in which students came from the same or comparable schools in the same or comparable districts, or used student-level covariates to postequate students on important demographic and/or achievement characteristics. The majority of these studies also were conducted with language-minority students of elementary

school age. Most were longitudinal, and children had transitioned out of the bilingual program by posttest. A few studies were 1-year studies of bilingual education, with posttests being administered before children had transitioned from native- to English-language instruction. These studies are included in this chapter because they shed some light on the development of literacy skills for these learners in different instructional conditions. Because the programs had not been completed by children in the study, however, they are of limited value for making claims about the overall effectiveness of bilingual education programs.

In addressing the inherent problem of selection bias, the studies of Huzar (1973) and Plante (1976) are particularly important, despite taking place a quarter of a century or more ago. Both were multiyear experiments for which, because of the use of random assignment, we 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 2 to 3 years. It is interesting that both used a model that would be unusual today—paired bilingual reading instruction provided 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 resembles the experience of Spanish-dominant students in two-way bilingual programs (see Calderón & Minaya-Rowe, 2003) more than typical transitional bilingual models, which delay English reading instruction to second or third grade.

Finally, with respect to language-minority students experiencing reading difficulties, Maldonado's (1994) study of language-minority students with learning disabilities found dramatically higher achievement gains for children transitioned over a 3-year period from Spanish to English than for those taught only in English. Although results reported for Maldonado (1994) are not internally consistent with respect to the point estimate of the treatment effect, both sets of reported results indicate a large, positive effect.

FRENCH IMMERSION STUDIES

As discussed earlier, although the studies conducted in Canada that examined the impact of French immersion programs are not directly relevant to the effectiveness of bilingual programs for language-minority learners learning the societal language, they are of value in gaining a broader understanding of the role of context in literacy development. Several French immersion studies (e.g., Barik & Swain, 1978; Genesee et al., 1976) have played an important role in debates about bilingual education. In these studies, English-speaking children (Anglophones) were taught entirely or primarily in French in the early elementary years. Rossell and Baker (1996) emphasize that these studies are examples of structured immersion, the approach favored in their review. However, Willig (1985) and other reviewers excluded them because the Canadian context differs significantly from that of the United States, as elaborated in the background section of the chapter. Moreover, in contrast to U.S. studies, the focus of the Canadian studies was primarily on whether French immersion hinders the English-language development of native English speakers. This would be analogous to determining whether English immersion hinders the development of Spanish in Spanish-speaking language-minority students.

Three studies (Barik & Swain, 1975, 1978; Barik, Swain, & Nwanunobi, 1977) met our methodological criteria for inclusion in this chapter. Each was conducted with children in the elementary school years.

Barik and Swain (1975) studied a French immersion program in Ottawa. One cohort of students was followed from kindergarten through second grade. One group of Anglophone children was taught entirely in French in kindergarten and first grade, with 60 minutes of daily English instruction in second grade, in comparison with Anglophone children taught only in English. The children were matched with respect to age, IQ, and school readiness measures administered in kindergarten. On a measure of English reading administered at the end of second grade, there were no differences between the groups. A second cohort of students was followed from kindergarten through first grade. All were Anglophone students, some in French immersion classes and some in regular English classes. At the end of first grade, the French immersion students scored significantly lower than the comparison group on all three English-language measures (word knowledge, word discrimination, and reading). It should be noted, however, that the French immersion students had received no instruction in English at this point. A third cohort included two groups of Anglophone students at the end of kindergarten. As with the other cohorts, one group was in French immersion and the other in English-only instruction. At the end of kindergarten, there were no reliable differences on either the school readiness or achievement test. The immersion group scored much higher than the comparison group on French comprehension. The comparison group had received 20 to 30 minutes a day of French as a second language.

In another evaluation of French immersion (see Barik & Swain, 1978; Barik, Swain, & Nwanunobi, 1977), the English-language performance of three cohorts of children in a bilingual program, ranging from third through sixth grades, was evaluated for 2 consecutive years in comparison with that of a cohort who received all instruction in English. The children in the French immersion program were instructed in French in mathematics, music, science, and French language arts for half of the day and were instructed in English in English language arts, physical education, and other content areas for the other half of the day. The comparison group came from a demographically similar school located near the school from which the intervention group students were selected. The comparison students were instructed only in English. For the sample that was followed from third through fourth grade, the children in the bilingual program had higher scores at fourth grade on measures of English reading comprehension and English vocabulary. For the sample followed from fourth through fifth grade, there were no differences between groups at fifth grade on measures of English reading. Similarly, for the group followed from fifth through sixth grade, there were no differences between the groups at sixth grade on measures of English reading.

The findings from these French immersion 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 an opportunity to acquire facility in a second language. The relevance of these findings to the U.S. situation is in (a) suggesting that similar second-language immersion programs, as well as two-way bilingual programs, for English-proficient

children are not likely to hinder English reading development; and (b) providing a better understanding of how context influences learning.

METHODOLOGICAL ISSUES

Research on language of instruction faces a number of inherent issues beyond those typical of other research on educational programs. We address these issues here, as well as briefly in chapter 13.

First, many of the studies reported on in this chapter failed to account for differences in the amount of time language-minority students instructed in a bilingual setting had to acquire English before being evaluated against children instructed only in English; the point at which students are evaluated in their second language has an impact on the study findings. For example, imagine that a bilingual program teaches Spanish-dominant English-language learners primarily in Spanish in Grades K to 2 and then gradually transitions them to English, completing the process by fourth grade. If this program is compared with an English-only program, at what grade level is it legitimate to assess the children in English? Clearly, a test in second grade may be meaningless because the bilingual program children have not been taught to read in English. At the end of third grade, the bilingual program students have been partially transitioned, but have they had enough time to become fully proficient? As a specific example, Saldate et al. (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 lower than the immersion group in English reading, although the differences were not statistically significant. A 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 make such a comparison fairly because in the bilingual program children would need a reasonable time to transfer their Spanish reading skills to English (see e.g., Hakuta, Butler, & Witt, 2000).

A longitudinal study by Gersten and Woodward (1995), not included in this chapter because both groups received Spanish instruction, sheds some light on this issue. This study was carried out with Spanish-dominant language-minority learners in 10 El Paso, Texas, elementary schools. Five schools used a paired bilingual model, in which all subjects were taught in English, but Spanish instruction was also provided each day—for 90 minutes in first grade, declining to 30 minutes in fourth grade. The other schools followed a transitional bilingual model, which involved mainly Spanish instruction, with 1 hour per day of ESOL instruction, with a gradual transition to English being completed only in fourth or fifth grade. The children were well matched demographically at entry into first grade and scored near zero on a measure of English-language proficiency. In Grades 4, 5, 6, and 7, students from the two groups were compared in English reading by using the Iowa Tests of Basic Skills (ITBS). On total reading, the paired bilingual students scored significantly higher than the transitional bilingual students in fourth grade, but the effects diminished in fifth grade and were very small in sixth

and seventh grades. Similar results were seen on tests of language and vocabulary. This pattern of results is probably due to the fact that, in fourth and fifth grades, the transitional bilingual students had not completed their transition to English; when they had done so, by sixth grade, their reading performance was nearly identical to that of the paired bilingual group. The overall effect size for differences was small (.07)

Other problems that characterize this research relate to selection bias. Children end up in transitional bilingual education or English immersion by processes that could have a significant impact on the outcomes, regardless of language of instruction. For example, Spanish-dominant students may be assigned to Spanish or English instruction within a school because of parental preferences in ways that have an impact on outcomes. Parents who select English programs may differ consistently from those who select Spanish programs. A parent who selects English may be less likely to be planning to return to a Spanish-speaking country, for example, or may feel more positive about assimilation. Likewise, a parent who selects Spanish may be from a home where little English is spoken. In addition, schools may assign individual children 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 language-minority students who are more dominant in their first than in their second language.

Further, bilingual programs are more likely to exist in schools with high proportions of English-language learners, and this is another potential source of bias. For example, Ramírez et al. (1991) found that schools using a late-exit bilingual approach had much higher proportions of English-language learners than early-exit bilingual schools, and English immersion schools had the smallest proportion of such learners. Regardless of the language of instruction, children in schools with high proportions of language-minority students, especially those from the same language background, are probably conversing less with native English speakers both in and out of school than might be the case in an integrated school that uses English for all students because its proportion of language-minority students is low. A related issue in some evaluations is that children in the bilingual program consist of only those who have not transitioned out of the bilingual program (and thus those who have taken longer to become proficient in English); these students are compared with those who have been instructed only in English, as well as those who have transitioned out of bilingual programs. The study by Danoff et al. (1978), for example, has been criticized for comparing children in transitional bilingual education programs with those who have transitioned out of these programs.

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

Finally, many studies do not provide sufficient detail about the interventions to demonstrate just what is working with these students. Moreover, no study we

reviewed collected fidelity data on the bilingual program or carefully assessed the nature and quality of the instruction provided to students.

OVERALL SUMMARY

In summary, there is no indication that bilingual instruction impedes academic achievement in either the native language or English, whether for language-minority students, students receiving heritage language instruction, or those enrolled in French immersion programs. Where differences were observed, on average they favored the students in a bilingual program. The meta-analytic results clearly suggest a positive effect for bilingual instruction that is moderate in size. This conclusion held up across the entire collection of studies and within the subset of studies that used random assignment of students to conditions. Supporting the argument for high-quality studies in this area, those studies in which there was random assignment to conditions (Covey, 1973; Huzar, 1973; Kaufman, 1968; Maldonado, 1994; Plante, 1976) found significant differences in favor of the students receiving native-language instruction, with effect sizes ranging from small (.01) to large (.77), exclusive of the very large effect in Maldonado (1994), and a significant average effect size across the collection of studies, regardless of which statistical model is assumed for the distribution of effect sizes (fixed effects model or random effects model).

What is also of interest and worthy of further research is that three of the studies (Huzar, 1973; Maldonado, 1994; Plante, 1976) whose results favored bilingual programs evaluated models that are a variation on the more common models of bilingual education. Each of these studies was conducted with children in the early elementary years, and one (Maldonado, 1994) with a specific sample of Spanish speakers receiving special education services for learning disabilities. Both Huzar (1973) and Plante (1976) used paired bilingual models in which children were taught reading in both English and Spanish daily, at different times of the day. In the study by Maldonado (1994), the children receiving bilingual special education were taught to read in Spanish for the first year, in Spanish and English in the second year, and in English in the third year—a more rapid transition than is typical of some transitional bilingual programs. As a group, these studies suggest an intriguing possibility: English-language learners may learn to read best if taught in both their native language and English from early in the process of formal schooling. Rather than confusing children, as some have feared, reading instruction in a familiar language may serve as a bridge to success in English because decoding, sound blending, and generic comprehension strategies clearly transfer between languages that use phonetic orthographies, such as Spanish, French, and English (see chap. 9, this volume; August, 2002; August & Hakuta, 1997; Fitzgerald, 1995a, 1995b; García, 2000).

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

In addition to the few randomized experiments included in this chapter, the majority of studies with language-minority students, as well as the heritage language and French immersion studies, used a matched design with experimental and control groups. Taken together, the findings from these studies suggest that there are no negative effects and, in many cases, positive effects of bilingual approaches to instruction.

As noted previously, research on language of instruction may suffer from the tendency for journals to publish only articles that find significant differences. Another form of bias is the selection of exemplary programs for research purposes (e.g., the collection of studies included in Campeau et al., 1975). Given that dissertations and technical reports are less likely to suffer from such bias, we included them in our review.

Overall, where differences between two instructional conditions were found in the studies reviewed, these differences typically favored the bilingual instruction condition. This is the case for studies conducted with students in both elementary and secondary schools, and with students possessing a range of abilities. For example, the results of the one study designed to evaluate bilingual instruction for a specific population—Spanish speakers receiving special education services—favored a bilingual approach for these learners. Moreover, children in the bilingual programs studied not only developed facility with English literacy to the same extent as their peers educated in English, but also developed literacy skills in their native language. Thus, they achieved the advantage of being bilingual and biliterate.

Because of the inherent methodological problems cited in this chapter, an adequate study comparing bilingual and monolingual approaches would randomly assign a large number of children to be taught in English or their native language; pretest them on outcomes of interest, as well as on language proficiency in their first and second languages; and 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 permit a fair comparison. In addition, researchers would carefully document the nature and quality of the instruction being provided. Unfortunately, only a few small studies of this kind have ever been conducted. As a result, the findings of studies that have compared bilingual and English-only approaches must continue to be interpreted with great caution.[9]

[9]IES is currently funding three evaluation studies employing experimental or quasi-experimental methods and will compare outcomes for students instructed in English only with those instructed with some use of the native language.

APPENDIX 14.A

TABLE 14.A.1
Evaluation Studies Cited in Other Reviews, But Not Included in This Review

Citation	*Reasons for Rejection From This Review*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
American Institutes for Research, 1975	Unavailable		X		
Ames & Bicks, 1978	No appropriate experimental group	X	X		
Ariza, 1988	Publication type: conference paper		X		
Bacon et al., 1982	No pretest measures		X	X	X
Balasubramonian et al., 1973	4 months between pre- and posttests		X		
Barclay, 1969	Unavailable		X		
Bates, 1970	No pretest measures, no literacy outcomes		X		
Becker & Gersten, 1982	Not an evaluation of bilingual programs				
Bruck et al., 1977	No adequate pretests		X		
Burkheimer et al., 1989	No appropriate control group		X		X
Carsrud & Curtis, 1979, 1980	No appropriate control group	X	X		
Ciriza, 1990	Unavailable		X		
Clerc et al., 1987	Unavailable		X		
Cohen, 1975	Publication type: book	X	X		
Cottrell, 1971	No appropriate control group		X		X

(Continued)

TABLE 14.A.1
(Continued)

Citation	*Reasons for Rejection From This Review*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
Curiel, 1979	Redundant with Curiel, 1980; no data while students in bilingual education				
Curiel et al., 1980	No pretest measures; no data while students were in bilingual education		X		
Danoff et al., 1977a, 1977b, 1978			X		
Day & Shapson, 1988	6-week duraction of bilingual instruction	X			
Educational Operations Concepts, 1991	Unavailable		X		X
El Paso ISD, 1987	No appropriate pretest measure		X		
El Paso ISD, 1990	Redundant with El Paso, 1992		X		
El Paso ISD, 1992	Comparison of two bilingual programs		X		
Elizondo de Weffer, 1972	No literacy measures		X		
Genesee, Lambert, & Tucker, 1979	No adequate pretests		X		
Genesee & Lambert, 1983	No appropriate control group		X		X

(Continued)

TABLE 14.A.1
(Continued)

Citation	*Reasons for Rejection From This Review*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
Genesee et al., 1989	No adequate pretests		X		X
Gersten, 1985	No appropriate control group		X		
Lambert & Tucker, 1972	Publication type: book		X		X
Layden, 1972	Only 10-week interval between pre- and posttests		X		
Legarreta, 1979	No literacy outcomes	X	X		
Lum, 1971	No literacy outcomes	X	X		
Malherbe, 1946	No appropriate control group		X		
Matthews, 1979	No specificity about native-language instruction		X		
McConnell, 1980a, 1980b	Prospective case study		X		
McSpadden, 1979	Unavailable	X	X		
McSpadden, 1980	Unavailable	X	X		
Medina & Escamilla, 1992	No literacy outcomes		X		
Meléndez, 1980	No appropriate experimental group		X		
Moore & Parr, 1978	Mixed Spanish- and English-dominant students		X		

(Continued)

TABLE 14.A.1
(Continued)

Citation	*Reasons for Rejection From This Review*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
Olesini, 1971	Unavailable	X			
Pena-Hughes & Solis, 1980	Unavailable	X	X		
Powers, 1978	No pretest measures		X	X	
Prewitt-Díaz, 1979	17-week interval between pre- and posttests		X		
Ramos et al., 1967	English as a foreign language		X		
Rossell, 1990	Book chapter		X	X	
Rothfarb et al., 1987	Both groups exposed to formal Spanish instruction		X	X	
Skoczylas, 1972	No literacy outcomes	X	X	X	
Stebbins et al., 1977	No specificity about native-language instruction	X	X	X	
Stern, 1975	Confounds with respect to instruction; no appropriate comparison groups	X	X		
Teschner, 1990	Unavailable	X	X		
Vásquez, 1990	No pretest controls		X		
Yap et al., 1988	No appropriate experimental groups		X		
Zirkel, 1972	No literacy outcomes	X	X		

APPENDIX 14.B: CODING SHEET FOR EVALUATION STUDIES

To determine study disposition, please answer each of the following questions:

1. The study	Yes	No
a. compared language-minority children.[10]	o	o
b. taught literacy.[11]	o	o
c. was in classes/programs that used some native language.	o	o
d. was against classes/programs that were taught in English.[12]	o	o

Note that the native-language group cannot be compared against tabled normative information.

2. English is the societal language (except in parts of Canada where French is the societal language).	o	o
3. If no to #1, the study is based in Canada and compared...		
a. English-dominant students acquiring literacy in French as a second language.	o	o
b. to English-dominant students learning mostly in English.	o	o
4. Random assignment to conditions was used.	o	o
5. If no to #4, a control or comparison group was used and there was some assessment of comparability prior to onset of the time interval over which the inference is being made (e.g., a pretest was used).[13]	o	o
6. The language-minority students in the sample are either at least 50% of the sample or the outcome data are disaggregated by language minority status (except for French immersion studies).	o	o
7. The interval between the pre- and posttests is at least 6 months.	o	o

Study is accepted if:

- Yes to all parts of Question 1 OR yes to all parts of Question 3 AND
- Yes for 2, AND
- Yes to 4 or 5, AND
- Yes to 6 and 7.

If study is ACCEPTED, answer Question 8

8. Serious confounds exist in the design of the research that prevent effects from being reasonably attributed to the treatment variables of interest.[14]

If yes, please explain:

[10]Language-minority students are students who come from a home where a language other than English is spoken. For the purposes of our work, we also include native Hawaiian children, Alaska natives, and American Indians even if the home language is not specified.

[11]Literacy includes reading as well as skills related to reading such as writing, vocabulary, and comprehension. Studies of oral language proficiency alone are not included.

[12]Note that some English immersion classes use small amounts of the native language to clarify concepts. This still constitutes an English-only class or program.

[13]Some studies consider tests administered at the end of the year as pretests. However, such tests are not considered as pretests for our coding purposes.

[14]For example, a study that compares two programs that both use some native language instruction (Carlisle & Beeman, 2000) would be excluded, as would a study (Curiel, 1980) that does not use random assignment or include pretest data or a description of the control group.

APPENDIX 14.C

TABLE 14.C.1
Evaluation Studies Included in the Present Review and Other Reviews

Study	*Characteristics*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
Alvarez, 1975	Elementary school; matched design		X		X
Campeau et al., 1975	Elementary school; matched design		X		X
Cohen et al., 1976	Elementary school; matched design		X		
Covey, 1973	Secondary school; random assignment	X	X	X	X
Danoff et al., 1978	Elementary school; matched design	X	X	X	
de La Garza & Medina, 1985	Elementary school; matched design		X		
Doebler & Mardis, 1980–81*	Elementary school; matched design				X
Huzar, 1973	Elementary school; random assignment	X	X	X	X

(*Continued*)

TABLE 14.C.1
(Continued)

Study	*Characteristics*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
Kaufman, 1968	Secondary school; random assignment	X	X	X	X
Lampman, 1973*	Elementary School; Matched Design		X		
Maldonado, 1977	Elementary school; matched design		X		
Maldonado, 1994	Elementary school; random assignment				X
Plante, 1976	Elementary school; random assignment				X
Ramírez, 1991	Elementary school; matched design		X	X	X
Saldate et al., 1985	Elementary school; matched design				
Valladolid, 1991	Elementary school; matched design		X		

TABLE 14.C.1
(Continued)

Study	*Characteristics*	*Willig, 1985*	*Rossell & Baker, 1996*	*Greene, 1997*	*Slavin & Cheung, 2004*
			Heritage Language Studies		
Morgan, 1971	Elementary school; matched design		X		X
			French Immersion Studies		
Barik & Swain, 1975*	Elementary school; matched design		X		X
Barik & Swain, 1978*	Elementary school; matched design		X		X
Barik et al., 1977*	Elementary school; matched design		X		X

* Not included in the meta-analysis.

APPENDIX 14.D: ADDITIONAL NOTES ON METHODOLOGY OF STUDIES CITED IN CHAPTER AND EFFECT SIZE CALCULATIONS

Alvarez, 1975

All students in the study were nonrepeaters and nontransfers. Information on attrition is not reported. Tabled effect sizes are on adjusted means and are not strictly comparable to effect sizes computed on unadjusted means. Also, the formula used is for analysis of variance (ANOVA) *F*, but the *F* reported is the *F* for groups in analysis of covariance (ANCOVA). The partial correlation is not reported for the covariate. The study reports unadjusted means and the covariate means, but no standard deviations are reported. The mean-square within is not reported for the ANOVA *F*, so it is not possible to estimate the standard deviation from the reported statistics. All effect sizes computed here are biased away from 0 for this reason; the standard deviation is underestimated because the covariate effect is included in the computation of *F*, but cannot be extracted from the effect size computation. Another concern is that the posttest scores appear to be grade equivalent scores.

Campeau et al., 1975

To obtain pretest effect sizes, standard deviations were taken from the posttest data for the same cohort, grade, and school district. If no standard deviations were reported for a grade/cohort/school, then the effect size was not estimated unless a standard deviation could be derived from reported statistics for that test measure from other studies in the pool of reviewed studies, or from another source using comparable samples. Specifically, we were able to estimate the variance for the IART Reading Total (RTT) measure in Grades 1 to 4 and extrapolate based on the trend there to Grade 5, insofar as the standard deviations were increasing about 5 points per grade. No effect sizes could be estimated for Santa Fe, New Mexcico, or for the VOC and Reading Comp measures. In the Corpus Christi sample, standard deviations were reported. For Alice and Houston, standard deviations were estimated based on the other studies in the pool as just described. The posttest standard deviation for the control group was assumed equal to that of the treatment group because it was not reported.

Cohen, 1976

Although the IART was used in each year, the levels are footnoted as Levels I, II, and III. To compute effect sizes, standard deviations were taken from other studies that used the IART at the same grades. We computed the average variance estimate for all studies using the IART RTT at a given grade, and we took the square root of the mean variance. These were not weighted by sample size. Standard deviations for the IART RTT were estimated for each grade based on other studies in the pool because Cohen did not report them (see Campeau).

Huzar, 1973

In Tables 4 and 5, the author reports results for separate bilingual classrooms that differed either by having one bilingual and one monolingual teacher or two bilingual teachers. Because there were two experimental classes at each grade, this information can be used to determine the magnitude of the intraclass

correlation (ICC) within the experimental condition. Assuming this ICC is consistent across first- and second-language classrooms (a nontrivial assumption) allows us to correct standard errors and point estimates for clustering. These tables are used to estimate the magnitude of the ICC in Grades 2 and 3. Within-class sample sizes are not given in the tables, but can be inferred from the standard errors of the means, which are given. Because random assignment was used, the posttest *d* was not adjusted.

Kaufman, 1968

The *F* statistics reported are ANCOVA *F*s, but information on the R^2 for covariates is not presented. Consequently, ANOVA *F* conversions are used to estimate the effect size, which tend to underestimate the standard deviation in the population. Thus, all reported effect sizes are biased away from 0 (i.e., are more positive or more negative than they would be if a measure of the standard deviation were available).

Maldonado, 1994

It appears that one or more errors are present in the reporting of the data. Effect sizes based on the means and standard deviations are most likely wrong (*d* = 7.007). If the standard deviations reported are actually standard errors, then the effect size is a more realistic 2.2 and compares reasonably well with an effect size based on the reported *t* statistic, 1.7. Values currently reported in Table 14.1 for standard deviations are those reported in the article multiplied by the square root of the sample size. The standard deviations reported seem low given the metric of the test (mean = 100, *SD* = 15), and it appears that the reported standard deviations could be standard errors, although they are clearly labeled as standard deviations. Computing *d* using the original means and taking the standard deviations to be standard errors gives an effect size of 2.2. If the pre- and postmeans have been reversed for the control group in the table, the reported *t* statistics are close if the standard deviations are also taken to be standard errors. They are much too small if the reported standard deviations are indeed standard deviations. Taking the posttest mean for the control group to be 69, rather than 63 as is reported in the table, the effect size is 1.70, close to that based on the reported *t* statistic. No adjustments completely reconcile the reported statistics with one another.

Maldonado, 1977

The analysis is for Grades 2 to 5 outcomes. Science Research Associates (SRA) forms were different in each grade. The reported regression tables use the group variable as the dependent variable and list the outcome and covariate as predictors. To obtain unadjusted effect size estimates, the regression model was solved for the bivariate correlations, and these were converted to *d* statistics using the *r* to *d* conversion in Lipsey and Wilson (2001). To determine the direction of the effect size, information was taken from the text.

To obtain the observed bivariate correlation between Group and the outcome, the regression models reported in the paper had to be reverse engineered. The table analyses use Group as the outcome and the covariate and outcome as the predictors. To obtain *r* between the desired outcome and Group, we employed the equation $R^2 = \beta_1 r_{yx1} + B_2 r_{yx2}$ from Pedhazur (1997). Because R_2 and the betas are reported in the table and R_2 for the covariate is given

separately, it is possible to solve the equation for r_{xz}, which can then be converted to *d* using the formula in Lipsey and Wilson (2001). In the text, we are told that the negative coefficient for fourth-grade mathematics indicates a mean difference in favor of the control group. Hence, it is assumed that positive betas on the outcome indicate an effect favoring the treatment group, whereas negative betas indicate an effect favoring the control group.

Morgan, 1971

Computed *d* is posttest *d* without correction for pretest. Groups were said to be equated on the pretest Metropolitan Achievement Test (MAT), but data were not provided.

Plante, 1976

Clustering of students in classrooms was ignored, and teachers were not randomly assigned to classrooms. Although ICCs at Grade 2 are small, this is not the case at Grade 3. The investigator also compared the groups on Metropolitan Readiness Test (MRT) and IQ, and groups were not statistically different. Tables 4 and 5 provide data on individual classes in the experimental group in Grades 2 and 3, respectively. These tables can be used to estimate the magnitude of the clustering effect in the bilingual classrooms, which was small in Grade 2 (.008), but large in Grade 3 (.78) because of the large mean difference between the two classrooms. Because random assignment was used, the posttest *d* was not adjusted.

Saldate et al., 1985

We imputed first-language standard deviation at pretest to be equal to second-language standard deviation at pretest. No standard deviation was given for first language at pretest.

APPENDIX 14.E

TABLE 14.E.1
Effect Sizes of Studies Included in the Meta-Analysis

Study Name	*Subgroup Within Study*	*Outcome*	*Time Point*	*Biling-ED N*	*English-Only N*	*Std Diff in Means*	*Std Err*	*Hedges' g*	*Standard Error*
Maldonado, 1994	Sample 1	Reading total	2	10	10	2.215	0.568	2.121	0.544
Saldate et al., 1985	Sample 1	Unknown	2	31	31	-0.287	0.255	-0.283	0.252
Saldate et al., 1985	Sample 1	Word reading	3	19	19	0.908	0.341	0.889	0.334
de la Garza, 1985	Sample 1	Reading comprehension	2	25	117	0.192	0.221	0.191	0.219
de la Garza, 1985	Sample 1	Reading comprehension	3	25	117	0.207	0.221	0.206	0.219
de la Garza, 1985	Sample 1	Reading vocabulary	2	24	118	0.496	0.226	0.494	0.225
de la Garza, 1985	Sample 1	Reading vocabulary	3	24	118	0.249	0.224	0.248	0.223
Ramirez et al., 1991	Sample 1	Reading total	1	67	139	0.178	0.149	0.177	0.148
Ramirez et al., 1991	Sample 1	Reading total	2	67	139	-0.258	0.149	-0.257	0.149
Ramirez et al., 1991	Sample 1	Reading total	3	67	139	0.154	0.149	0.154	0.148
Ramirez et al., 1991	Sample 2	Reading total	1	252	194	0.095	0.096	0.095	0.095
Ramirez et al., 1991	Sample 2	Reading total	2	252	194	-0.100	0.096	-0.099	0.095
Ramirez et al., 1991	Sample 2	Reading total	3	252	194	0.017	0.096	0.017	0.095
Ramirez et al., 1991	Sample 3	Reading total	1	170	194	0.080	0.105	0.080	0.105
Ramirez et al., 1991	Sample 3	Reading total	2	170	194	-0.276	0.106	-0.275	0.105
Ramirez et al., 1991	Sample 3	Reading total	3	170	194	-0.067	0.105	-0.067	0.105
Valladolid, 1991	Sample 1	Reading total	4	50	57	-0.610	0.198	-0.605	0.197
Valladolid, 1991	Sample 1	Reading total	5	50	57	-0.541	0.197	-0.538	0.196
Alvarez, 1975	Sample 1	Reading comprehension	2	51	26	-0.188	0.241	-0.186	0.239
Alvarez, 1975	Sample 1	Reading vocabulary	2	51	26	0.163	0.241	0.162	0.239

(Continued)

TABLE 14.E.1
(Continued)

Study Name	*Subgroup Within Study*	*Outcome*	*Time Point*	*Biling-ED N*	*English-Only N*	*Std Diff in Means*	*Std Err*	*Hedges' $g^{(i)}$*	*Standard Error*
Alvarez, 1975	Sample 2	Reading comprehension	2	39	31	-0.257	0.242	-0.254	0.239
Alvarez, 1975	Sample 2	Reading vocabulary	2	39	31	0.072	0.241	0.071	0.238
Campeau et al., 1975	Sample 2	Reading total	1	104	27	1.839	0.244	1.828	0.243
Campeau et al., 1975	Sample 2	Reading total	2	94	21	0.924	0.249	0.918	0.247
Campeau et al., 1975	Sample 2	Reading total	3	75	22	-0.241	0.243	-0.239	0.241
Campeau et al., 1975	Sample 3	Reading total	1	106	19	1.401	0.264	1.393	0.263
Campeau et al., 1975	Sample 3	Reading total	2	95	35	1.434	0.217	1.426	0.216
Campeau et al., 1975	Sample 3	Reading total	3	101	29	0.761	0.216	0.757	0.215
Campeau et al., 1975	Sample 3	Reading total	4	75	20	0.271	0.252	0.269	0.250
Campeau et al., 1975	Sample 5	Reading total	1	125	46	2.643	0.224	2.631	0.223
Campeau et al., 1975	Sample 5	Reading total	2	97	57	0.406	0.168	0.404	0.168
Campeau et al., 1975	Sample 5	Reading total	3	85	63	0.882	0.174	0.877	0.173
Campeau et al., 1975	Sample 6	Reading total	1	119	100	0.243	0.136	0.242	0.136
Campeau et al., 1975	Sample 6	Reading total	2	205	93	0.355	0.126	0.354	0.126
Campeau et al., 1975	Sample 6	Reading total	3	79	80	0.389	0.160	0.387	0.159
Campeau et al., 1975	Sample 7	Reading total	1	146	60	0.857	0.159	0.854	0.158
Campeau et al., 1975	Sample 7	Reading total	2	161	53	0.668	0.162	0.665	0.161
Campeau et al., 1975	Sample 7	Reading total	3	218	83	0.123	0.129	0.123	0.129
Campeau et al., 1975	Sample 7	Reading total	4	98	88	0.760	0.152	0.757	0.151
Campeau et al., 1975	Sample 8	Reading total	1	145	45	0.457	0.172	0.455	0.172
Campeau et al., 1975	Sample 8	Reading total	2	155	53	0.541	0.161	0.539	0.161
Campeau et al., 1975	Sample 8	Reading total	3	146	62	0.272	0.152	0.271	0.152
Campeau et al., 1975	Sample 8	Reading total	4	151	58	0.390	0.156	0.389	0.155
Cohen et al., 1976	Sample 1	Reading total	4	14	11	-0.180	0.404	-0.174	0.390
Cohen et al., 1976	Sample 1	Reading Total	5	7	7	-0.220	0.536	-0.206	0.502
Cohen et al., 1976	Sample 2	Reading total	3	12	9	-1.200	0.478	-1.152	0.459

(*Continued*)

TABLE 14.E.1
(Continued)

Study Name	*Subgroup Within Study*	*Outcome*	*Time Point*	*Biling-ED N*	*English-Only N*	*Std Diff in Means*	*Std Err*	*Hedges' s*[1]	*Standard Error*
Cohen et al., 1976	Sample 2	Reading total	4	7	7	-1.145	0.577	-1.072	0.540
Cohen et al., 1976	Sample 3	Reading total	2	7	9	-1.690	0.586	-1.598	0.554
Cohen et al., 1976	Sample 3	Reading total	3	7	7	-1.809	0.635	-1.694	0.594
Danoff et al., 1977	Sample 1	Reading total	2	722	297	-0.262	0.069	-0.262	0.069
Danoff et al., 1977	Sample 1	Reading total	3	905	469	-0.265	0.057	-0.265	0.057
Danoff et al., 1977	Sample 1	Reading total	4	941	515	0.097	0.055	0.097	0.055
Danoff et al., 1977	Sample 1	Reading total	5	731	144	-0.287	0.091	-0.287	0.091
Huzar, 1973	Sample 1	Reading total	6	341	68	-0.424	0.133	-0.423	0.133
Huzar, 1973	Sample 1	Reading total	2	41	40	0.014	0.222	0.014	0.220
Kaufman, 1968	Sample 1	Paragraph meaning	3	43	36	0.313	0.227	0.310	0.225
Kaufman, 1968	Sample 1	Paragraph meaning	7	41	31	0.048	0.238	0.048	0.235
Kaufman, 1968	Sample 1	Word meaning	8	31	19	0.115	0.292	0.113	0.287
Kaufman, 1968	Sample 1	Word meaning	7	41	31	0.220	0.239	0.217	0.236
Kaufman, 1968	Sample 2	Paragraph meaning	8	31	19	0.311	0.293	0.306	0.288
Kaufman, 1968	Sample 2	Word meaning	7	20	25	0.478	0.304	0.470	0.299
Maldonado, 1977	Sample 1	Reading total	2	20	25	0.039	0.300	0.038	0.295
Maldonado, 1977	Sample 1	Reading total	3	47	79	0.360	0.186	0.358	0.184
Maldonado, 1977	Sample 1	Reading total	4	47	79	0.506	0.187	0.503	0.186
Maldonado, 1977	Sample 1	Reading total	5	47	79	0.475	0.187	0.473	0.186
Plante, 1976	Sample 1	Reading total	2	15	10	0.378	0.186	0.376	0.185
Plante, 1976	Sample 1	Reading total	3	16	12	0.801	0.424	0.775	0.410
Covey, 1973	Sample 1	Reading total	9	89	84	0.272	0.384	0.264	0.372
Morgan, 1971	Sample 1	Paragraph reading	1	93	100	0.661	0.156	0.658	0.156
Morgan, 1971	Sample 1	Word reading	1	93	100	0.255	0.145	0.254	0.144
						0.374	0.145	0.372	0.145