The Big Picture: A Meta-Analysis of Program Effectiveness Research on English Language Learners

KELLIE ROLSTAD, KATE MAHONEY, and GENE V. GLASS

This article presents a meta-analysis of program effectiveness research on English language learners. The study includes a corpus of 17 studies conducted since Willig's earlier meta-analysis and uses Glass, McGaw, and Smith's strategy of including as many studies as possible in the analysis rather than excluding some on the basis of a priori "study quality" criteria. It is shown that bilingual education is consistently superior to all-English approaches, and that developmental bilingual education programs are superior to transitional bilingual education programs. The meta-analysis of studies controlling for English-language-learner status indicates a positive effect for bilingual education of .23 standard deviations, with outcome measures in the native language showing a positive effect of .86 standard deviations. It is concluded that bilingual education programs are effective in promoting academic achievement, and that sound educational policy should permit and even encourage the development and implementation of bilingual education programs.

Keywords: *bilingual education; meta-analysis; evaluation; English-only education; English-as-a-second-language education; structured English immersion*

THE QUESTION OF HOW BEST TO EDUCATE English language learners (ELLs) enrolled in U.S. schools is an important one. Kindler (2002) reported federally collected demographic statistics estimating that 4,584,946

EDUCATIONAL POLICY, Vol. 19 No. 4, September 2005 572-594
DOI: 10.1177/0895904805278067

ELLs were enrolled in U.S. public schools in the 2000 to 2001 school year, an approximate increase of 32% over reported 1997 to 1998 enrollments. According to Kindler, California enrolled the largest number of ELLs in public schools (1,511,646), followed by Texas (570,022), Florida (254,517), New York (239,097), Illinois (140,528), and Arizona (135,248); the Marshall Islands, Micronesia, Palau, and Puerto Rico reported 100% of their public school students to be ELLs. P. R. Campbell (1994) projected the proportional rate of increase in the ELL population from 1995 to 2020 to be 96%, compared to an expected increase of 22% for native English speakers. Given these facts, and the legal responsibility placed on schools to provide effective programs for ELL students (*Lau v. Nichols*, 1974), it comes as no surprise that considerable attention has been given to the question of how schools best serve these students.

Several states, including California (Proposition 227), Arizona (Proposition 203), and Massachusetts (Question 2), have passed ballot initiatives that restrict the types of educational method and program that may be used to instruct ELLs. While these laws have substantially restricted bilingual education in California and Massachusetts, an aggressive implementation of the law in Arizona by the current Superintendent of Public Instruction has made bilingual education essentially impossible in that state (Mahoney, Thompson, & MacSwan, 2004).[1] At the federal level, the Bilingual Education Act of 1968, which had been repeatedly reauthorized, was repealed concurrently with the passage of the No Child Left Behind Act and replaced with the English Acquisition Act. As its name implies, the English Acquisition Act emphasizes the acquisition of English rather than dual-language instruction, and it imposes new accountability measures on schools, pressuring them to emphasize rapid transition to English-only instruction, a typical focus of mandated English-only instructional programs (Crawford, 2004).

Because the education of students who are immigrants is closely tied to issues of nationalism, immigration, and the politics of multilingualism, the debate over how best to serve ELL students has often been clouded by politics (Petrovic, 1997). As Rossell and Baker (1996) commented

> This field is so ideologically charged that no one is immune from ideological bias or preconceived notions. As a result, those attempting to make policy recommendations from the research must carefully read each study and draw their own conclusions. This does not guarantee that such conclusions will be free from bias, only that they will be free from someone else's bias. (pp. 25-26)

Although we agree that research may never be entirely free from researchers' perspectives, we believe that a fair and reasonable description of the available

evidence is possible, and that a properly conducted meta-analysis will help provide a factual description of program effects across a large range of available studies.

In the current study, we present a meta-analysis of studies comparing effects of instructional programs for ELL students in an effort to clarify "the big picture" in this debate. Our approach differs from previously conducted literature reviews in that it includes many studies not reviewed previously, and we did not exclude studies a priori based on design quality. Although our corpus and methodological approach differ from those of previous researchers, our conclusions are consistent with most of the major reviews conducted to date. We find an advantage for approaches that provide instruction in the students' first language and conclude that state and federal policies restricting or discouraging the use of the native language in programs for ELL students cannot be justified by a reasonable consideration of the evidence.

PREVIOUS RESEARCH SYNTHESES

Hundreds of methods comparison studies and program evaluations have attempted to ascertain whether bilingual or English-only instruction best serves ELLs. Although it is beyond the scope of the current study to provide an exhaustive review of previously published research syntheses on this topic (Baker & de Kanter, 1981; Demmert & Towner, 2003; Dulay & Burt, 1978; Engle, 1975; Epstein, 1977; Greene, 1998; Holland, 1986; McField, 2002; Peterson, 1976; Rossell & Baker, 1996; Rossell & Ross, 1986; Rotberg, 1982; Slavin & Cheung, 2003; Troike, 1978; Willig, 1985; Yates & Ortiz, 1983; Zappert & Cruz, 1977; Zirkel, 1972), we briefly review five studies that have been widely cited in the policy arena. These reports fall into two broad categories: narrative reviews, which include Baker and de Kanter (1981), Rossell and Baker (1996), and Slavin and Cheung (2003); and previous meta-analyses, by Greene (1998) and Willig (1985).

The Narrative Reviews

Baker and de Kanter (1981) reported that they considered more than 300 studies that compared effects of language of instruction, from which they selected 28, published between 1968 and 1980, as methodologically sound studies. They rejected studies in which (a) their research question was not addressed; (b) students were not randomly assigned to the treatment and comparison groups, or nothing was done to control for possible initial differences between the groups; (c) appropriate statistical tests were not used to demonstrate program effects; (d) the norm-referenced design was used, comparing student growth against test norms; (e) gains during the school year

were examined without a control group; or (f) grade-equivalent scores, strongly criticized by Horst and colleagues (1980), were used.

Baker and de Kanter (1981) were interested in comparing transitional bilingual education (TBE) to three alternatives: submersion, English as a second language (ESL), and structured immersion (SI). They defined TBE as a program in which subject matter is taught in the children's home language until their English is strong enough for them to participate in an all-English classroom, with the use of the native language gradually phasing out and the use of English gradually phasing in. The ESL approach was defined as a program in which children are placed in regular (not sheltered) English-only classes for most of the day, and provided with concentrated instruction aimed at teaching English as a second language during part of the day. SI, on the other hand, provides a specially structured version of the regular curriculum so that students may acquire the language of instruction while simultaneously learning content; in SI, the teacher knows the children's home language; however, it is rarely or never spoken by the teacher in the classroom.

Baker and de Kanter painted a mixed picture. Studies included in the review varied in their conclusions, with some findings indicating that TBE was better, as good as, or no different from SI, others that ESL was better or as good as SI and TBE. The authors nonetheless made more favorable comments about the success of SI than they did about other programs. With regard to the comparison between TBE and SI, two studies were included in the review. One of these, conducted in the Philippines to teach English as a second language, found no difference between SI and TBE (Ramos, Aguilar, & Sibayan, 1967). The other (Peña-Hughes & Solis, 1980), according to Baker and de Kanter, found that students in SI outperformed students in TBE. However, apart from the short duration of the program (9 months) in the study, the SI label was not consistent with Baker and de Kanter's program descriptions, as Willig (1985) pointed out. In this McAllen, Texas, kindergarten program, students spent a portion of every instructional day in native language instruction (Peña-Hughes & Solis, 1980), a characteristic of TBE according to Baker and de Kanter's definitions. Therefore, according to Baker and de Kanter's program definitions, this program would properly be labeled TBE rather than SI, and its results should be interpreted as favoring TBE.

Baker and de Kanter ended their study with the recommendation that government policy should be flexible in the domain of program selection and development:

> We conclude that it is very hard to say what kind of program will succeed in a particular school. Hence it seems that the only appropriate Federal policy is to allow schools

to develop instructional programs that suit the unique needs and circumstances of their students. (chap. 4, p. 6)

In a study intended to update the research of Baker and de Kanter (1981) and other reviews, Rossell and Baker (1996) found 72 studies to be methodologically sound, based on the following selection criteria, similar to those used by Baker and de Kanter (1981): The included studies (a) were true experiments in which students were randomly assigned to treatment and control groups, or had nonrandom assignment that either matched students in the treatment and comparison groups on factors that influence achievement or statistically controlled for them; (b) included a comparison group of ELL students of the same ethnicity and similar language background; (c) used outcome measures in English using normal curve equivalents (NCEs), raw scores, scale scores, or percentiles, but not grade equivalents; and (d) involved no additional educational treatments, or controlled for these treatments if present.

Rossell and Baker (1996) included as SI programs those that "typically include at least 30-60 minutes a day of native language arts beginning sometime in the early elementary years" (p. 10). This conception of SI is a departure from that proposed in Baker and de Kanter (1981), where SI is distinguished from bilingual instruction in that "the home language (L1) is never spoken by the teacher and subject area instruction is given in the second language from the beginning" (chap. 1, p. 2). If we consider bilingual education to be simply "the use of the native language to instruct limited English-speaking children" (Rossell & Baker, 1996, p. 1), then it appears that the authors' SI program description overlaps in significant respects with their bilingual education program description. These imprecise definitions make it difficult to know whether a program labeled *immersion* in a study was not actually a bilingual education program for the purposes of Rossell and Baker's review. Like Baker and de Kanter (1981), Rossell and Baker concluded that there remains "no consistent research support for transitional bilingual education as a superior instructional practice for improving the English language achievement of limited English proficient children" (p. 19).

Krashen (1996), Greene (1998) and Slavin and Cheung (2003) noted that many of the studies included in the Rossell and Baker study do not conform to their own selection criteria. For example, some studies included in the Rossell and Baker review did not include an appropriate comparison group. Burkheimer, Conger, Dunteman, Elliott and Mowbay (1989) and Gersten (1985) compared students' performance to statistical estimates of where they should have been performing. Rossell (1990) compared an ethnically diverse group of immersion students (48% Asian) to an ethnically homogeneous

group of bilingual students (100% Hispanic). Rossell and Baker (1996) included this study in their review despite claiming that only studies in which a comparison group of ELL students of the same ethnicity and similar language background were included in their narrative review. Furthermore, Slavin and Cheung (2003) noted that the Rossell and Baker review, like the Baker and de Kanter review before it, included Canadian studies of French immersion (e.g., Genesee & Lambert, 1983; Lambert & Tucker, 1972) in which immersion was compared to monolingual English instruction of children who already knew English. This instructional model is very different from structured immersion as used in the United States. Moreover, Slavin and Cheung (2003) pointed out that Rossell and Baker assigned multiple "votes" to studies that were published in multiple forms (Curiel, 1979; Curiel, Stenning, & Cooper-Stenning, 1980; El Paso Independent School District Office for Research and Evaluation, 1987, 1990, 1992), even though only one experiment was collectively reported. It should be noted that the studies assigned multiple votes presented results that Rossell and Baker (1996) interpreted as favoring immersion.

It is also important to note that Baker and de Kanter (1981) defined SI as a program in which the teacher "understands the home language (L1), and students can address the teacher in the home language (L1); the immersion teacher, however, replies in the second language (L2)" (chap. 1, p. 2). The stipulation that immersion teachers know the home language of students was also presented in Rossell and Baker (1996, p. 10). This factor distinguished SI in these narrative reviews from the version of the SI model mandated in California, Arizona, and Massachusetts, where ballot initiatives have imposed English immersion programs on all ELLs in the state. Because teachers in the mandated version of SI do not typically know the home language of their students, and are not required to know it, whatever evidence one might consider persuasive for the effectiveness of SI from these narrative reviews cannot be generalized to the legislative context of these states.

Baker and de Kanter's (1981) criteria for accepting studies relied on what they believed to be "general agreement in the scientific literature on what constitutes good study design" (chap. 1, p. 4) as outlined, they said, in such work as D. T. Campbell and Stanley (1963). Rossell and Baker (1996) relied on similar conventions based on similar authority. Slavin (1986) similarly included such considerations in his best-evidence approach, a version of the narrative review approach that, Slavin and Cheung (2003) remarked, additionally involves the systematic inclusion criteria and effect size computations typical of meta-analysis (Cooper, 1998; Glass, 1976; Glass, McGaw, & Smith, 1981) whenever such calculations are possible.

Slavin and Cheung (2003) recently completed a best evidence review of studies focused on methods of teaching reading to ELL students, comparing the practice of teaching ELLs to read in their native language first (a bilingual education strategy) with that of teaching them to read in English first (an immersion strategy). Following a broad search for all studies involving ELL students, assisted in part by outside organizations, Slavin and Cheung selected studies according to the following criteria: (a) the studies compared children taught reading in bilingual classes to those taught in English immersion classes; (b) 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; (c) the participants were ELLs in elementary or secondary schools in English-speaking countries; (d) the dependent variables included quantitative measures of English-reading performance, such as standardized tests and informal reading inventories; and (e) the treatment duration lasted at least one school year. Slavin and Cheung identified 16 studies, published between 1971 and 2000, that met these criteria.

Slavin and Cheung's review concluded that on balance the evidence favors bilingual approaches, especially paired bilingual strategies that teach reading in the native language and English at the same time. Most of the studies they found to be methodologically acceptable favored bilingual approaches over immersion approaches; although some found no difference, none significantly favored immersion programs.

The Meta-Analyses

Willig's (1985) and Greene's (1998) meta-analyses have provided the best published sources of integrated evidence thus far; however, both meta-analyses focused primarily on studies done before 1985. The current study updates the corpus of studies to be included in the meta-analysis, a statistical procedure specifically designed to summarize research findings to come to some general conclusions regarding effects or outcomes of a given treatment, project, or program (Glass et al., 1981).

In response to Baker and de Kanter's (1981) narrative review, Willig (1985) conducted a meta-analysis to determine if their conclusions could be sustained using similar statistical procedures. Rather than using the full corpus of studies on this topic for which meta-analysis is possible, Willig imposed still stricter selection criteria, requiring that studies focus on K-12 students in U.S. schools. As a result, Baker and de Kanter's 28 studies dropped to 23 in Willig's review. Although Baker and de Kanter's review framed their findings as supportive of local flexibility and emphasized that exclusive reliance on bilingual programs was not warranted, Willig (1985)

found "positive effects for bilingual programs . . . for all major academic areas" (p. 297).

Greene (1998) similarly produced a meta-analysis of the studies included in Rossell and Baker's (1996) narrative review, again imposing additional selection criteria, narrowing the corpus significantly to only 11 studies. Like Willig, Greene found positive effects for bilingual education:

> Despite the relatively small number of studies, the strength and consistency of these results, especially from the highest quality randomized experiments, increases confidence in the conclusion that bilingual programs are effective at increasing standardized test scores measured in English. (p. 5)

Thus, the general conclusions of Slavin and Cheung (2003) are consistent with Willig (1985) and Greene (1998). Their conclusions are at odds with Baker and de Kanter (1981) and Rossell and Baker (1996) in the sense that the latter concluded that the evidence on this topic is inconclusive. While the reports by Slavin and Cheung (2003), Willig (1985), and Greene (1998) did a better job of interpreting the results of the studies included in Baker and de Kanter (1981) and Rossell and Baker (1996) because of their more accurate identification and categorization of program models, all of these studies shared a practice of dubious utility: the a priori preselection of studies based on design features. As we discuss in the next section, a better approach to research synthesis involves meta-analysis in which the widest possible net is used to include relevant studies, subsequently narrowing the focus based on empirically detected effects of specific variables within the analysis.

Methodological Differences in Research Synthesis

There are two substantially different positions on the question of a priori selection of studies in research synthesis: Glass's (1976) original advice to cast the widest net in beginning a meta-analysis, and Slavin's (1986) argument that research synthesis should proceed on the basis of best evidence, a view consistent with the analyses of program effectiveness for ELL students reviewed above.

A disadvantage of the best-evidence approach is that the reviewer has great latitude in assessing how important any particular study is and, thus, imposes personal preferences on what is included (apparent, for instance, in the uneven application of selection criteria and program definitions in Baker reviews discussed previously (Baker & de Kanter, 1982; Rossell & Baker, 1996). Another weakness is the arbitrariness of the selection criteria. For instance, Slavin (1986) insists that random assignment of children to various programs be one condition for inclusion. Indeed, by far the greatest source of

experimental invalidity in educational experiments arises from the experimenter's inability to randomize influences at the level of classrooms and schools. However, random assignment of students to groups achieves only partial experimental control when intact classrooms themselves are not randomly assigned to the experimental conditions—thus leaving uncontrolled a host of factors that operate at the classroom level, such as teachers' expertise, time of day, and similar factors. Hence, Slavin's insistence on random assignment of students to treatments is not as crucial to experimental validity as some might think. The best evidence may not be as good as a great deal of evidence arbitrarily excluded by Slavin's approach.

When a sample of so-called methodologically acceptable studies has been assembled, the narrative approach uses vote counting of the results of individual studies to report the overall synthesis of the research literature. However, because vote counting does not systematically consider effect size, we do not know how much better one group did over another in any individual study.

Glass's approach, by contrast, advocated that researchers include as many studies as possible. As Glass et al. (1981) put it

> The goal of the meta-analyst should be to provide an accurate, impartial, quantitative description of the findings in a population of studies on a particular topic. . . . No survey would be considered valid if a sizable subset (or stratum) of the population was not represented in the cumulative results. Neither should a meta-analysis be considered complete if a subset of its population is omitted. (p. 64)

The current study differs from previous research syntheses on this topic in three respects. First, it provides an update to the corpus by including in the analysis only studies published in 1985 or later. Second, it provides comparisons not only for TBE and English-only approaches but also for developmental bilingual education[2] (DBE) as well. Third, our approach includes as many of these studies as possible in the meta-analysis and does not apply what we see as arbitrary, a priori selection criteria. This permits us to probe more deeply into the distribution of study results to understand why some studies may find a stronger advantage for a particular program than another, a matter we return to in our discussion of the results.

METHOD

Selecting the Studies

In an effort to focus on recent research, we limited our search to studies completed after Willig's (1985) meta-analysis. Thus, we searched ERIC, PsychInfo, and Dissertation Abstracts for evaluation studies addressing

programs for language minority students with publication dates of 1985 or later. More than 300 studies were identified and reviewed.

Studies were included in the current meta-analysis according to the following selection criteria: They (a) involved K-12 language minority students (who were not enrolled in special education classes), (b) included statistical details needed to perform the meta-analysis, and (c) provided a description of the treatment and comparison programs. As a consequence, we could not include studies that did not use comparative research methods, involved a treatment other than a program for ELLs, confounded other treatments with the treatment of interest, reported too little data, or did not focus on program effectiveness.

Seventeen studies meeting these criteria were identified: Burnham-Massey (1990), Carlisle (1989), Carter and Chatfield (1986), de la Garza and Medina (1985), Gersten (1985), Gersten and Woodward (1995), Gersten et al. (1992), Lindholm (1991), Medina and Escamilla (1992), Medina et al. (1985), Medrano (1986, 1988), Ramirez et al. (1990), Rossell and Baker (1996), Rothfarb, Ariza, and Urrutia (1987), Saldate et al. (1985), and Texas Education Agency (1988).

Coding the Studies

When studies were identified, selected characteristics were coded and given quantitative descriptions. Broad categories of coded variables included study identification, characteristics of program, characteristics of students, characteristics of teachers, characteristics of research design, and effect size variables, as shown in Table 1. Because program labels for ELL students are often oversimplified or misleading, special caution was taken to code program type according to the actual description provided in the study's text.

Calculating Effect Size

The preferred formula for estimating effect size when integrating studies that compare a relatively new treatment with what might be called a traditional or control treatment is the difference between the mean of the new treatment group and the traditional treatment group on the final outcome measure, divided by the standard deviation of the traditional treatment group (Glass et al., 1981). Conceptually, we maintained the preferred formula; however, in the current study, we are not really comparing a new treatment to a traditional treatment in the literal sense of new and traditional because those terms, drawn from true experimental research design, cannot be properly applied to the quasi-experimental design commonly used in education research. To remain consistent and because we had a variety of program types to compare, we used a first comparison group and a second comparison group where the

Table 1
Coded Characteristics of the Studies

Study identification	Author's last name
	Year of publication
	Study identification number
	Publication form
Characteristics of program	Bilingual program type
	Use of native language
	Source of L1 support
	Model of L1 support
	Criteria used for LEP classification
	Length of time program continues in years
	L1 support used for content areas
Characteristics of students	Average grade level
	Percentage female
	Percentage male
	SES
	Ethnicity
	First language
Characteristics of teachers	Credentialed in bilingual education
	Proficient in student's language
	Years of experience teaching
Characteristics of research design	Type of group assignments
	Type of teacher assignments
	Control for SES
	Internal validity
	Number of comparisons in this study
Outcome measure characteristics	Sample size
	Mean
	Standard deviation
	Score form
	Instrument used for outcome measure
	Language of outcome measure
	Academic domain
	Source of means
	Calculation of effect size

Note: L1 = first language; LEP = limited English proficiency; SES = socioeconomic status.

first comparison group was always more aligned with bilingual education pedagogy. In some cases, probit transformations were used to calculate effect size. A probit transformation uses dichotomous data (e.g., above grade level vs. below grade level) to derive a standardized mean difference. It is based on the very simple assumption that if 84% of a normally distributed Group A is above some point, and 50% of a normally distributed Group B is above that

same point, then the means of Groups A and B must differ by one standard deviation.

For longitudinal or multiyear studies, an effect size was calculated for each year and each grade level. The first comparison group in every effect-size calculation was the comparison group implementing a program more aligned with bilingual education pedagogy. DBE was considered to represent the program most aligned with bilingual education followed, in order, by TBE, English as a second language and/or structured English immersion (ESL and/or SEI), and English-Only[1] (EO^1) for limited English proficient (LEP) students and English-Only[2] (EO^2) for non-limited English proficient (non-LEP) students.

RESULTS

Effect Sizes by Individual Studies

There is a wide range of variability in program, grade, sample size, and outcome measures. Please note the range of program comparisons (Table 2). We can confidently assert that the experimental group was aligned more with bilingual education pedagogy; however, the program type and comparison group vary from study to study. All 17 studies used outcome measures derived from standardized tests; however, the instrument and the content area vary widely.

It is important to distinguish instances in which researchers made comparisons between two groups of ELLs from those in which comparisons were made between a group of ELLs and a group of native English-speaking students. There are many plausible reasons why comparing ELLs to native English-speaking students will yield a higher effect size for the native English-speaking group, all of which are unrelated to true achievement differences. For example, norming bias and construct irrelevant variance resulting from differences in language proficiency give an unfair advantage to native English-speaking children, unrelated to program effects. We see that the average effect size is highly influenced by whether the researcher chose to compare ELLs to other ELLs or to native speakers of English. In this meta-analysis, ELLs are compared to native speakers of English in six instances (TBE vs. EO^2, 5 times; DBE vs. EO^2, 1 time). There are also 13 instances where ELLs are compared to other ELLs: TBE vs. ESL (7 instances), DBE vs. EO^1 (3 instances), TBE vs. EO^1 (1 instance), DBE vs. ESL (1 instance), and DBE vs. TBE (1 instance).

When coded, the 17 studies yielded a total of 156 instances where two different bilingual programs were compared on one or more outcome measures,

(text continued on p. 588)

Table 2
Comparisons of Effect Size (ES) by Study

Study	N *of ES*	M *ES*	*SD of ES*[a]
Burnham-Massey, 1990			
Grades 7-8			
Range of *n*'s for TBE: 36 to 115			
Range of *n*'s for EO[2]: 36 to 115			
TBE vs. EO[2]			
Reading	3	–.04	.07
Mathematics	3	.24	.14
Language	3	.16	.25
Carlisle, 1989			
Grade 4, 6			
Range of *n*'s for TBE: 23			
Range of *n*'s for EO[1]: 19			
Range of *n*'s for EO[2]: 22			
TBE vs. EO[1]			
Writing-rhetorical effectiveness	1	.82	
Writing-overall quality	1	1.38	
Writing-productivity	1	.60	
Writing-syntactic maturity	1	1.06	
Writing-error frequency	1	.50	
TBE vs. EO[2]			
Writing-rhetorical effectiveness	1	-2.45	
Writing-overall quality	1	-8.25	
Writing-productivity	1	.18	
Writing-syntactic maturity	1	.24	
Writing-error frequency	1	1.01	
Carter and Chatfield, 1986			
Grades 4-6			
Range of *n*'s for DBE: 26 to 33			
Range of *n*'s for EO[2]: 14 to 47			
DBE vs. EO[2]			
Reading	3	.32	.24
Mathematics	3	–.27	1.06
Language	3	–.60	1.54
de la Garza and Medina, 1985			
Grades 1-3			
Range of *n*'s for TBE: 24 to 25			
Range of *n*'s for EO[2]: 116 to 118			
TBE vs. EO[2]			
Reading vocabulary	3	.15	.38
Reading comprehension	3	.17	.06
Mathematics computation	3	–.02	.15
Mathematics concepts	3	–.02	.14

(continued)

Table 2 (continued)

Study	N *of ES*	M *ES*	*SD of ES*[a]
Gersten, 1985			
Grade 2			
Range of *n*'s for TBE: 7 to 9			
Range of *n*'s for ESL: 12 to 16			
TBE vs. ESL			
Reading	1	-1.53	
Mathematics	1	–.70	
Language	1	-1.44	
Gersten, Woodward, and Schneider, 1992			
Grades 4-6			
Range of *n*'s for TBE: 114 to 119			
Range of *n*'s for ESL: 109 to 114			
TBE vs. ESL			
Reading	4	–.17	.12
Language	4	–.35	.26
Mathematics	4	.00	.17
Gersten and Woodward, 1995			
Grades 4-7			
Range of *n*'s for TBE: 117			
Range of *n*'s for ESL: 111			
TBE vs. ESL			
Reading	4	–.15	.13
Language	4	–.33	.22
Vocabulary	3	–.15	.12
Lindholm, 1991			
Grades 2-3			
Range of *n*'s for DBE: 18 to 34			
Range of *n*'s for EO[1]: 20 to21			
DBE vs. EO[1]			
Reading	1	–.59	
Language	2	–.14	.57
Medina and Escamilla, 1992			
Grades K-2			
Range of *n*'s for DBE: 138			
Range of *n*'s for TBE: 123			
DBE vs. TBE			
Language-oral, native	2	.64	.74
Language-oral, English	1	.11	

(continued)

Table 2 (continued)

Study	N *of ES*	M *ES*	*SD of ES*[a]
Medina, Saldate, and Mishra, 1985			
Grades 6, 8, 12			
Range of *n*'s for DBE: 19			
Range of *n*'s for EO[1]: 24 to 25			
DBE vs. EO[1]			
Metropolitan Achievement Test			
Total mathematics	2	–.32	.16
Problem solving	2	–.24	.13
Concepts	2	–.34	.25
Computation	2	–.13	.53
Total reading	2	–.21	.08
Reading	2	–.30	.28
Word knowledge	2	–.10	.10
California Achievement Test			
Total mathematics	1	–.20	
Concepts/application	1	–.11	
Computation	1	–.27	
Total reading	1	–.63	
Comprehension	1	–.57	
Vocabulary	1	–.41	
Medrano, 1986			
Grades 1, 6			
Range of *n*'s for TBE: 179			
Range of *n*'s for EO[2]: 108			
TBE vs. EO[2]			
Reading	2	–.18	.13
Mathematics	2	.10	.24
Medrano, 1988			
Grades 1, 3			
Range of *n*'s for TBE: 172			
Range of *n*'s for EO[2]: 102			
TBE vs. EO[2]			
Reading	1	.10	
Mathematics	1	.60	
Ramirez, Yuen, Ramey, Pasta, and Billings, 1991			
Grades 1-3			
Range of *n*'s for DBE: 97 to 197			
Range of *n*'s for TBE: 108 to 193			
Range of *n*'s for ESL: 81 to 226			
DBE vs. ESL			
Mathematics	3	.26	.22
Language	3	–.43	–.97
Reading	3	.37	.21

(continued)

Table 2 (continued)

Study	N *of ES*	M *ES*	*SD of ES*[a]
TBE vs. ESL			
Mathematics	3	.11	.10
Language	3	–.17	.17
Reading	3	.01	.16
Rossell, 1990			
Grades K-12			
Range of *n*'s for TBE: 250			
Range of *n*'s for ESL: 326			
TBE vs. ESL			
Oral language	2	.36	.23
Rothfarb, Ariza, and Urrutia, 1987			
Grades 1-2			
Range of *n*'s for TBE: 34 to 70			
Range of *n*'s for ESL: 33 to 49			
TBE vs. ESL			
Tests in English			
Mathematics	4	.13	.11
Language	2	.28	
Social studies	4	.20	.13
Science	4	.09	.18
Tests in Spanish			
Mathematics	4	.11	.14
Language	2	.10	
Social studies	4	.23	.22
Science	4	.16	.11
Saldate, Mishra, and Medina, 1985			
Grades 2-3			
Range of *n*'s for DBE: 31			
Range of *n*'s for EO[1]: 31			
DBE vs. EO[1]			
Tests in English			
Total achievement[a]	1	–.29	
Reading	1	1.47	
Spelling	1	.50	
Arithmetic	1	1.16	
Tests in Spanish			
Total achievement	1	.46	
Reading	1	2.31[b]	
Spelling	1	3.03	
Arithmetic	1	1.16	

(continued)

Table 2 (continued)

Study	N *of ES*	M *ES*	*SD of ES*[a]
Texas Education Agency, 1988			
Grades 1, 3, 5, 7, 9			
Range of *n*'s for TBE: approximately 135,000			
Range of *n*'s for ESL: approximately 135,000			
TBE vs. ESL			
Tests in English			
Mathematics	4	–.03	.02
Reading	4	–.06	.13
Tests in Spanish			
Mathematics	2	.33	.06
Reading	2	.78	.09

Note: TBE = transitional bilingual education; DBE = developmental bilingual education; ESL = English as a Second Language; EO^1 = English-Only instruction for children with limited English proficiency; EO^2 = English-Only instruction for non-LEP students.
a. Reading, spelling, and arithmetic are not constituents of the total achievement.
b. This effect size was calculated with the treatment group's standard deviation.
c. CTBS scores in Rossell (1990) were excluded from the analysis because scores were not available for all students in the sample.

and effect sizes were calculated for all 156. Table 2 lists the 17 studies and their mean effect sizes, and the standard deviation for each outcome variable represented in the study. Four of the 17 studies report outcome measures in the students' native language in addition to English (Medina & Escamilla, 1992; Rothfarb et al., 1987; Saldate et al., 1985; and Texas Education Agency, 1988).

A positive effect size indicates that the bilingual program group fared better than the comparison group, whereas a negative effect size indicates that the comparison group fared better. The magnitude of an effect size indicates the between-group difference in units of the standard deviation of the control group. For example, de la Garza and Medina (1985) compared TBE to EO for non-limited English proficient students in Grades 1 through 3. Their study showed the size of the sample for the TBE group to be about one fifth the size of the EO group. The mean effect size for reading vocabulary was calculated as .15. This indicates that ELL students exposed to TBE scored about one sixth of a standard deviation higher than the EO group made of non-ELL students.

Inspecting the results of the studies shown in Table 2, one notes that Carlisle (1989) reported a dramatic difference between TBE and EO^2 (language majority students) students in two writing measures. While comparisons of ELL students to language majority students is problematic on independent grounds (as we discuss directly), the reported effect sizes for these measures were wildly inconsistent with those reported in other studies

Table 3
Combining Effect Sizes (ES) by Grouping

Grouping	N *of ES*	M *ES*	SD *of ES*
All outcome measures	67	.08	.67
Reading (in English)	16	–.06	.61
Math (in English)	15	.08	.42
All outcomes in native language	11	.86	.96
Without Gersten studies	58	.17	.64
Without Medrano studies	64	.07	.69
Without Medina studies	44	.17	.76
Language minority students vs. Language majority students	14	.05	.28
Language minority students vs. Language minority students	22	.23	.97
All TBE studies	35	–.01	.45
All DBE studies	30	.18	.86

Note: TBE = transitional bilingual education; DBE = developmental bilingual education.

making similar comparisons. This observation led us to treat the measures reported in Carlisle as outliers, and to exclude them from the calculations presented in Table 3.

Table 3 gives overall effect size results for our corpus of studies from a variety of perspectives. Most notably, 14 effect sizes from the current study are derived from comparisons made between ELLs and EO^2s, who are not limited in their English proficiency, yielding an overall positive effect for bilingual education of .05. However, it has been widely noted that ELL status entails many more disadvantages in addition to limited proficiency in English (August & Hakuta, 1998), so that studies in which comparisons are made between ELLs and non-ELLs inevitably confound the treatment effect with numerous other unknown factors. Indeed, as shown in Table 3, studies that compare two groups of ELLs, and therefore control for ELL status, render a much higher positive effect of .23 for bilingual education.

To investigate the possibility of a researcher effect, authors who contributed more than 1 of the 17 studies were removed from the overall effect size results in Table 3. When Gersten's studies are removed from the meta-analysis, the overall effect size increases from .08 to .17. Considering Gersten contributes 9/67 of the effect sizes in the meta-analysis, this has a significant impact on the results and may show signs of a researcher effect. When results were grouped by type of bilingual program, DBE studies showed a much higher effect size (.18) than TBE (–.01).

When outcomes were measured in the native language rather than English, the positive effect for bilingual education over alternative approaches

increases dramatically, to .86. The positive effects of bilingual instruction may be more readily detected by Spanish-medium measures of academic achievement in the absence of limited proficiency in English as a source of measurement error in such tests (Thompson, DiCerbo, Mahoney, & MacSwan, 2002).

Finally, we note that the current meta-analysis reveals not only that bilingual education is superior to all-English approaches such as ESL or SI but also that programs designed to develop children's academic use of both languages (DBE) are superior to programs that aim to use children's home language to transition them to all-English instruction (TBE).

Our results are similar to those of Willig (1985), who reported a positive effect for bilingual education in reading (.20) and math (.18), measured in English, and for all outcomes in the native language (.69). Greene (1998) similarly found a positive effect for bilingual education in reading (.21) and math (.12) and in all outcomes measured in the native language (.74).

CONCLUSIONS

Empirical evidence considered here indicates that bilingual education is more beneficial for ELL students than all-English approaches such as ESL and SI. Moreover, students in long-term DBE programs performed better than students in short-term TBE programs. As expected, the effect is particularly strong in studies that controlled for ELL status.

It seems clear from the current study and from previous meta-analyses (Greene, 1998; Willig, 1985) that bilingual education is superior to English-only approaches in increasing measures of students' academic achievement in English and in the native language. In addition, well-conducted narrative synthesis, in which careful attention is given to an even application of selection criteria and program definitions (e.g., Slavin & Cheung, 2003), also conclude that bilingual education approaches are superior to all-English approaches for ELL students.

In view of these results, current policies implemented in California, Arizona, and Massachusetts, which ban or greatly discourage the use of the native language for instructional purposes, cannot be justified. Furthermore, the tendency of federal policies embedded in the No Child Left Behind (NCLB) Act to emulate these restrictive policies by emphasizing rapid transition to English are also ill advised. Instead, a rational educational policy, unencumbered by politics and ideology, should at least permit, and at best encourage, the development and implementation of bilingual education approaches in all U.S. schools serving ELLs.

NOTES

1. In Colorado, voters rejected a similar initiative (see Escamilla, Shannon, Carlos, & Garcia, 2003 for discussion).

2. Programs designed to develop academic use of both languages for children with a minority language.

REFERENCES

August, D., & Hakuta, K. (1998). *Improving schooling for language-minority children: A research agenda*. Washington, DC: National Academy Press.

Baker, K., & de Kanter, A. A. (1981). *Effectiveness of bilingual education: A review of the literature* (Final draft report). Washington, DC: Department of Education, Office of Planning, Budget, and Evaluation.

Burkheimer, G., Jr., Conger, A., Dunteman, G., Elliott, B., & Mowbray, K. (1989). *Effectiveness of services for language minority limited English proficient students*. Raleigh-Durham, NC: Research Triangle Institute.

Burnham-Massey, M. (1990). Effects of bilingual instruction on English academic achievement of LEP students. *Reading Improvement*, *27*, 129-132.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Campbell, P. R. (1994). *Population projections for states by age, race, and sex: 1993 to 2020*. Washington, DC: U.S. Bureau of the Census.

Carlisle, R. S. (1989). The writing of Anglo and Hispanic elementary school students in bilingual, submersion, and regular programs. *Studies in Second Language Acquisition*, *11*, 257-280.

Carter, T. P., & Chatfield, M.L. (1986). Effective bilingual schools: Implications for policy and practice. *American Journal of Education, 95*(1), 200-232.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Crawford, J. (2004). *Educating English learners: Language diversity in the classroom* (5th ed.). Los Angeles: Bilingual Education Services.

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

Curiel, H., Stenning, W., & Cooper-Stenning, P. (1980). Achieved reading level, self-esteem, and grades as related to length of exposure to bilingual education. *Hispanic Journal of Behavioral Sciences*, *2*(4), 389-400.

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

Demmert, W. G., Jr., & Towner, J. C. (2003). *A review of the research literature on the influences of culturally based education on the academic performance of Native American students*. Portland, OR: Northwest Regional Educational Laboratory.

Dulay, H. C., & Burt, M. K. (1978). From research to method in bilingual education. In J. E. Alatis (Ed.), *Georgetown University Roundtable on Language and Linguistics* (pp. 551-575). Washington, DC: Georgetown University Press.

Escamilla, K., Shannon, S., Carlos, S. & García, J. (2003). Breaking the code: Colorado's defeat of the anti-bilingual education initiative (Amendment 31). *Bilingual Research Journal*, *27*(1), 357-382.

Engle, P. (1975). The use of the vernacular language in education. In *Bilingual Education* (pp. 1-33). Washington, DC: Center for Applied Linguistics.

Epstein, N. (1977). *Language, ethnicity and the schools: Policy alternatives for bilingual-bicultural education.* Washington, DC: Institute for Educational Leadership.

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

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

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

Gersten, R., Woodward, J., & Schneider, S. (1992). *Bilingual immersion: A longitudinal evaluation of the El Paso program*. Washington, DC: READ Institute. (ERIC Document Reproduction Service No. ED389162)

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

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

Greene, J. P. (1998). *A meta-analysis of the effectiveness of bilingual education*. Claremont, CA: Thomas Rivera Policy Institute.

Holland, R. (1986). *Bilingual education: Recent evaluations of local school district programs and related research on second-language learning*. Washington, DC: Congressional Research Service.

Horst, D. P., Johnson, D. M., Nava, H. G., Douglas, D. E., Friendly, L. D., & Roberts, A. O. H. (1980). *Prototype evaluation manual. In An evaluation of project information packages (PIPs) as used for the diffusion of bilingual projects* (Vol. 3). Mountain View, CA: RMC Corporation.

Kindler, A. (2002). *Survey of the states' limited English proficient students and available educational programs, 2000-2001 summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Krashen, S. (1996). *Under attack: The case against bilingual education.* Culver City, CA: Language Education Associates.

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

Lau v. Nichols, 414 U.S. 563 (1974).

Lindholm, K. J. (1991). Theoretical assumptions and empirical evidence for academic achievement in two languages. *Hispanic Journal of Behavioral Sciences*, *13*(1), 3-17.

Mahoney, K. S., Thompson, M. S., & MacSwan, J. (2004). The condition of English language learners in Arizona: 2004. In A. Molnar (Ed.), *The condition of pre-K-12 education in Arizona: 2004* (pp. 3.1-3.27). Tempe: Education Policy Studies Laboratory, Arizona State University.

McField, G. (2002). *Does program quality matter? A meta-analysis of select bilingual education studies*. Unpublished doctoral dissertation, University of Southern California, Los Angeles.

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

Medina, M., Jr., Saldate, M., IV, & Mishra, S. (1985). The sustaining effects of bilingual instruction: A follow-up study. *Journal of Instructional Psychology*, *12*(3), 132-139.

Medrano, M. F. (1986). Evaluating the long-term effects of a bilingual education program: A study of Mexican students. *Journal of Educational Equity and Leadership*, *6*, 129-138.

Medrano, M. F. (1988). The effects of bilingual education on reading and mathematics achievement: A longitudinal case study. *Equity and Excellence*, *23*(4), 17-19.

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

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

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

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

Peterson, M. (1976). *Assessment of the status of bilingual vocational training: Vol. 3. Review of the literature*. Albuquerque, NM: Kirschner Associates.

Petrovic, J. E. (1997). Balkanization, bilingualism, and comparisons of language situations at home and abroad. *Bilingual Research Journal*, *21*(2-3), 233-254.

Ramirez, J. D., Yuen, S. D., Ramey, D. R., Pasta, D. J., & Billings, D. (1991). *Final report: Longitudinal study of immersion strategy, early-exit and late-exit transitional bilingual education programs for language-minority children*. San Mateo, CA: Aguirre International. (ERIC Document Reproduction Service No. ED330216)

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

Rossell, C. H. (1990). The effectiveness of educational alternatives for limited-English proficient children. In G. Imoff (Ed.), *Learning in two languages* (pp. 71-121). New Brunswick, NJ: Transatlantic Publishers.

Rossell, C. H., & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English*, *30*(1), 7-74.

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

Rotberg, I. (1982). Federal policy in bilingual education. *American Education*, *52*, 30-40.

Rothfarb, S., Ariza, M. J., & Urrutia, R. (1987). *Evaluation of the Bilingual Curriculum Content (BCC) Pilot Project: A three year study* (Final report. Miami, FL: Dade County Public Schools, Office of Educational Accountability. (ERIC Document Reproduction Service No. ED300382)

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

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

Slavin, R. E., & Cheung, A. (2003). *Effective reading programs for English language learners: A best-evidence synthesis*. Baltimore: Johns Hopkins University Center for Research on the Education of Students Placed At Risk (CRESPAR).

Texas Education Agency. (1988). *Bilingual/ESL education: Program evaluation report*. Austin: Texas Education Agency. (ERIC Document Reproduction Service No. ED305821)

Thompson, M. S., DiCerbo, K., Mahoney, K. S., & MacSwan, J. (2002). ¿Éxito en California? A validity critique of language program evaluations and analysis of English learner test scores. *Education Policy Analysis Archives*, *10*(7). Available at http://epaa.asu.edu/epaa/v10n7/

Troike, R. (1978). Research evidence for the effectiveness of bilingual education. *NABE Journal*, *3*, 13-14.

Willig, A. C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research*, *55*(3), 269-318.

Yates, J., & Ortiz, A. (1983). Baker and de Kanter review: Inappropriate conclusions on the efficacy of bilingual programs. *NABE Journal*, *7*, 75-84.

Zappert, L., & Cruz, B. (1977). *Bilingual education: An appraisal of empirical research*. Berkeley, CA: Berkeley Unified School District.

Zirkel, P.-A. (1972). *An evaluation of the effectiveness of selected experimental bilingual education programs in Connecticut.* Unpublished doctoral dissertation, University of Connecticut, West Hartford.

Kellie Rolstad is assistant professor of language and literacy and early childhood education at Arizona State University. Her research interests include early childhood bilingualism, language diversity, bilingual education programs and theory, and two-way bilingual immersion. Her work has appeared in the Bilingual Research Journal, Bilingual Review, Teachers College Record, *and* Hispanic Journal of Behavioral Science, *among others, and she has served as a visitor in the Graduate School of Education at Harvard University and UCLA.*

Kate Mahoney is assistant professor in English as a Second Language (ESL) at SUNY Fredonia. She coordinates the M.Ed. Program concentration in ESL. Previously, she worked as a mathematics teacher for English language learners (ELLs). Her current research addresses uses of achievement and language proficiency test scores for ELLs, meta-analysis of program evaluations for students with a minority language, and the evaluation of policies and practices concerning ELLs in Arizona. Her work has appeared in Education Policy Analysis Archives *and* Bilingual Research Journal.

Gene V. Glass is Regents' Professor of Education Policy Studies and professor of psychology in education at Arizona State University. His Ph.D. was awarded by the University of Wisconsin, Madison, in educational psychology with a minor in statistics. From 1967 to 1986, he was on the faculty of the School of Education at the University of Colorado, Boulder. From 1997 to 2001, he served as associate dean for research in the Arizona State University College of Education. Trained originally in statistics, his interests broadened to include psychotherapy research, evaluation methodology, and policy analysis. In 1975, he was elected president of the American Educational Research Association. He served as editor of the Review of Educational Research *(1968-70), editor for* Methodology of the Psychological Bulletin *(1978-80), and coeditor of the* American Educational Research Journal *(1983-86). His work on meta-analysis of psychotherapy outcomes (with M. L. Smith) was named as one of the "Forty Studies that Changed Psychology" in the book of the same name by Roger R. Hock (1999). In recent years, his efforts have centered on the creation of new avenues for communications among researchers and the public. He was the founding editor of two electronic journals,* Education Policy Analysis Archives *and* Education Review *and is executive editor of the* International Journal of Education & the Arts.