Weighing the Evidence: A Meta-Analysis of Bilingual Education in Arizona

Kellie Rolstad, Kate S. Mahoney, and Gene V. Glass
Arizona State University

Abstract

This article reviews the current policy context in the state of Arizona for program options for English language learners and produces a meta-analysis of studies on the effectiveness of bilingual education that have been conducted in the state in or after 1985. The study presents an analysis of a sample of evaluation studies (N = 4), which demonstrates a positive effect for bilingual education on all measures, both in English and the native language of English language learners, when compared to English-only instructional alternatives. We conclude that current state policy is at odds with the best synthesis of the empirical evidence, and we recommend that current policy mandating English-only and forbidding bilingual education be abandoned in favor of program choices made at the level of the local community.

Introduction

Approximately 135,248 English language learners (ELLs) were enrolled in Arizona public schools in the 2000–2001 school year, making up about 15% of the state's total student enrollment; in proportion of ELLs, Arizona ranks third highest, behind only California and New Mexico (Kindler, 2002). In 2000, program options available to districts and parents to serve Arizona's ELL students were significantly changed with the passage of Proposition 203, a voter-approved initiative prohibiting bilingual education programs in the state. Proposition 203 ended the local flexibility in program options by repealing Article 3.1 of the Arizona Revised Statutes, replacing it with a requirement that all ELLs in the state be taught using Structured English Immersion (SEI) (Mahoney, Thompson, & MacSwan, 2004). Table 1 lists programs offered in

Table 1

Program Placements Reported to the Arizona Department of Education for English Language Learners Prior to the Implementation of Proposition 203 (Academic Year 1998–1999)

Program	Number of students	Percentage
Transitional bilingual education program Grades K–6	18,175	13
Secondary bilingual education program Grades 7–12	3,239	2
Bilingual/bicultural education program Grades K–12	23,505	16
English as a Second Language program Grades K–12	89,972	63
Individual education program	7,413	5
Individual education program, parental request	1,442	1
Total	143,746	100

Note. From Arizona Department of Education (2000).

the state prior to the passage of the initiative and the reported number of enrolled students. Table 2 outlines the Arizona Department of Education's language program categories for ELLs, with definitions of program models, as commonly understood in the literature (for review, see C. Baker, 2001; Crawford, 2004).

Proposition 203 was modeled after California's anti–bilingual education initiative, Proposition 227, which was passed in 1998. With the passage of Arizona's Proposition 203, districts scrambled to implement the new law but suffered from a lack of information, time, personnel, and resources to do so. For example, structured immersion, as defined by Baker and de Kanter (1981), requires that teachers understand the language of the children receiving structured immersion (SI), that teachers be trained in immersion methods, and that special curricula and materials be used. Many, if not most, Arizona districts implemented Proposition 203 without qualified bilingual teachers who had been trained in immersion methods and without adequate curricula and materials.

Table 2

Description of Arizona Department of Education's Language Program Categories for ELL Students in Arizona

Dates	Program	Description
Before 2000	Secondary bilingual Grades 7–12	Provides a portion of instruction in children's native language to help them keep up in school subjects while they learn English in programs designed for second-language learners. Programs additionally offer sheltered subject matter instruction (that is, second-language instruction that is "sheltered" from input beyond student's English comprehension).
Before 2000	Bilingual/bicultural Grades K–12	Otherwise known as maintenance or developmental bilingual education (MBE or DBE). These programs provide continued development in two languages.
Before 2000	English as a Second Language (ESL)	Involves language-sensitive content instruction for ELLs. Essentially all instruction is provided in English but with the curriculum and presentation designed for second-language learners. Also a component of bilingual programs.
Before 2000	Individual Education Plan (IEP)	Where appropriate programs are not available, an individual plan is designed to meet the child's language needs. Post 2000, this is no longer an allowable program option.
Before and after 2000	Transitional bilingual Grades K–6	Provides a portion of instruction in children's native language to help them keep up in school subjects while they learn English in programs designed for second-language learners. Students are transitioned to all-English instruction when their English and academic achievement are deemed sufficiently strong to allow full participation in an all-English instructional setting. After 2001, requires waiver for participation.

Table 2, cont.,

Description of Arizona Department of Education's Language Program Categories for ELL Students in Arizona

Dates	Program	Description
After 2000	Dual language	Also known as bilingual immersion and categorized as a type of DBE program, this program model works as an integrated approach in which monolingual English-speaking children are grouped with monolingual minority-language (e.g., Spanish) speaking children to learn each other's language and work academically in both languages. Requires waiver for language-minority participants, outlawing the model by definition in Arizona, since monolingual language-minority children are prohibited from non-English instruction.
After 2000	Structured English immersion (SEI)	Also known as sheltered English immersion, this model provides nearly all classroom instruction in English but with the curriculum and presentation designed for second-language learners. According to the literature, though not enforced in Arizona, the model requires that teachers know the home language of students.
After 2000	Mainstream English	English language learners receive no special language services. Appropriate, with monitoring, for fluent English proficient (FEP) students only.

Like the California initiative, the Arizona initiative permits waivers from SEI programs for some students. The law allows parents to submit waivers for children younger than 10 years of age who "already know English," further defined as possession of "good English language skills, as measured by oral evaluation or standardized tests of English vocabulary, comprehension, reading, and writing, in which the child scores approximately at or above the state average for his grade level or at or above the 5th grade average, whichever

is lower" (Arizona Revised Statutes, 2000). However, because the state grade level average on English oral language assessments has not been determined for students in Arizona, some districts asked test publishers for estimated averages while others estimated an average based on their own district testing data. However, in February 2003, the newly elected superintendent of public instruction, Tom Horne, who had run for office on a promise to "enforce" Proposition 203, issued guidelines that had the effect of altering the waiver requirement. Horne insisted that the "passing score" designated by the publisher—essentially arbitrary in nature (Glass, 1978)—rather than the grade-level average, serve as the minimum requirement for a waiver from English only classes. A state attorney general's opinion raised questions about the requirement, so Horne submitted average test scores supplied by the publisher for native speakers of English, not ELL students, and insisted that these serve as the waiver standard. A State Board of Education meeting left this construal of the law in place, and most bilingual programs were subsequently dismantled. These changes have imposed the most restrictive context in any state in the nation for native-language education, leaving parents who seek bilingual instruction for their children with no alternative to SEI. (See Mahoney et al. for further discussion.)

If it were the case that SEI were superior to bilingual approaches in educating ELLs, Horne's modifications of Proposition 203's requirements might be viewed as justified. However, a fair and reasonable consideration of the available evidence reveals that bilingual education programs have been more effective at raising students' test scores than all-English programs in Arizona, as we will show in this study. Striving to be fair and reasonable entails the use of research tools that rely as little as possible on subjective judgments. The statistical procedure of meta-analysis was specifically designed to summarize research findings regarding effects or outcomes of a given treatment, to provide clarity and increase objectivity when complexity threatens to obscure program outcomes (Glass, 1976; Glass, McGaw, & Smith, 1981). Although we focus on Arizona studies for present purposes, we believe that the results of meta-analysis as applied to the effects of bilingual education generalize to other U.S. contexts as well.[1]

Literature Review

In reviewing the relevant literature, it is important to understand that prior to the adoption of bilingual education programs, Arizona had a longstanding tradition of using English-only education. Known as English "1C courses," English-only instruction was mandated by Arizona law in 1919, and required that "all schools shall be conducted in English." In districts with large numbers of ELLs, these students received English vocabulary lessons in a low-level, simplified curriculum. Many of the students in 1C classes were

over age and often remained in 1C for several years before dropping out of school, never receiving the opportunity to transition to age-appropriate subject matter instruction. The 1C courses remained the only option for ELL students until 1965, when some bilingual programs were introduced (see Sheridan, 1986). Even then, a limit was placed on the number of bilingual programs that were permitted (Sacken & Medina, 1990), and English-only programs flourished.

Research Synthesis in Bilingual Education

Educational research generally, and bilingual education research specifically, is a complex undertaking that must attempt to come to terms with a dizzying array of variables. In order to understand how language of instruction affects language and subject matter learning among students who do not know English, for example, the researcher must consider myriad details that can independently affect outcomes, including students' prior knowledge of English and prior school experiences, whether students are literate in their first language, whether the teacher understands the language of the students, the teacher's years and type of teaching experience, what supports are available to aid students in both comprehension and production of English, use of visual aids and context clues in the classroom, length and intensity of interactions with peer language models, level and types of support available to students outside of class time, and many other potential influences.

The lack of consistency in program labels and definitions nationwide creates a thorny obstacle to research synthesis. A program labeled "English immersion" may provide several hours of native language instruction per day, while a program labeled "bilingual education" may provide no native-language instruction at all but rather a bilingual classroom aide for occasional translation support. A researcher interested in determining the effectiveness of a given program must often rely on guesswork when insufficient detail is provided in program evaluations, and different information may be provided depending on the needs and interests of various evaluators.

In addition, effectiveness itself is defined differently by different researchers. For some, a bilingual approach is determined to be superior to a monolingual approach if student achievement is similar to that attained by monolingual students, simply because bilingualism and biliteracy are additionally provided at no cost. This perspective follows from the view that bilingualism and biliteracy are potentially valuable as personal, cultural, economic, and globalizing enrichments. However, for others, a bilingual approach is regarded as superior only if its results, ultimately measured by achievement tests, are greater than those of ELL students in English-only programs. These two very distinct definitions of effectiveness influence the interpretation of program success.

An early attempt at research synthesis was made by Keith Baker and Adriana de Kanter (1981), who produced a narrative review of research conducted nationally on the effectiveness of bilingual education. At the time, Keith Baker was responsible for overseeing evaluation studies of language-minority education programs for the U.S. Department of Education. His position gave him great authority over whether and how findings of the federally funded studies were disseminated. In addition to his own narrative review, a major research synthesis encompassing national data and showing positive outcomes for bilingual education over all-English approaches was conducted during Baker's tenure. The study—a meta-analysis by Okada, Besel, Glass, Montoya-Tannatt, and Bachelor (1982)—revealed that children who received bilingual instructional support progressed at nearly twice the national norm in reading, math, and English language arts, with the strongest effects found in the early grades. Moreover, progress was greater when teachers were bilingual rather than monolingual speakers of English. Unfortunately, the report of the Okada and colleagues study was never released by Keith Baker's office.

In Baker's own narrative review with de Kanter (1981), released a year before the study by Okada and colleagues (1982) was completed, it was concluded that although bilingual instruction was found to be effective and superior to structured immersion in many cases, *exclusive reliance* on bilingual education was not justified. Baker and de Kanter called for local control and decision making regarding the education of ELLs:

> 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. (p. 17) (emphasis added)

As Secada (1987) pointed out, Baker and de Kanter's concerns regarding *exclusive reliance* and *federal policy* on bilingual education became moot soon thereafter, as no policy existed, then or ever, that mandated exclusive reliance on bilingual approaches. In subsequent decades, Baker and de Kanter's review has been widely, and mistakenly, cited as favoring SEI over bilingual education.

A striking, fundamental flaw in Baker and de Kanter's (1981) review is their inclusion of Canadian studies of French immersion programs. Baker and de Kanter's inclusion of these studies of structured *French* immersion in Canada, typically a 6-year or longer program, as evidence to support SEI, a short-term (not typically to exceed 1 year) *English* immersion program, has been widely criticized, for good reason, as the academic experiences of minority-language children are not comparable to the experiences of majority-language children (Dolson, 1985; Greene, 1998; Hernández-Chávez, 1984; Krashen, 1996; Malherbe, 1978; Secada, 1987; Slavin & Cheung, 2003; Tucker, 1980). French immersion programs for children who already speak English

begin with little or no use of the children's native language (English) but gradually increase in native-language use until a roughly 50–50 balance of English and French use is attained, typically by fifth or sixth grade. This is strikingly different from SEI in the United States, where programs begin with little or no native language use and end with none at all. Not surprisingly, the Canadian model produces students who are bilingual and biliterate, while SEI in the United States typically produces students who are proficient and literate only in English. This difference was not considered by Baker and de Kanter.

It should also be noted that Baker and de Kanter (1981) acknowledge as a necessary component of SEI that the teacher be bilingual; that 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)" (Chapter 1, p. 2). The stipulation that immersion teachers know the home language of students was also presented in a subsequent review by Rossell and Baker (1996), discussed later in the present article. The ability to understand the students in their care enables teachers to check for students' comprehension of classroom instruction and to understand students' questions in the classroom. Unfortunately, this crucial requirement of SEI, that teachers understand students' native language in order to provide SEI, is not clearly stated in laws such as Proposition 203 and remains largely unaddressed in district-level discussions of teacher qualifications and in selection of teachers to provide SEI.

In addition to issues of selection and interpretation, another fundamental matter in conducting research synthesis concerns the choice of narrative review versus meta-analysis, a choice not available until the latter was introduced (Glass, 1976). Narrative reviews such as Baker and de Kanter's (1981) have faced harsh criticism within many fields of research. For instance, Hunt (1997) argues against the narrative research review, making the following point:

> Although it offers a handy list of items in a particular area of research, it does little to integrate or cumulate them. Some reviews do offer more combinatory conclusions, but not methodically or rigorously; a recent critique of fifty medical review articles said that most summarized the pertinent findings in an unsystematic, subjective, and "armchair" fashion. (p. 7)

Hunt goes on to cite still harsher criticism of narrative reviews from prominent medical meta-analysts, Joseph Lau and Thomas Chalmers (1993), who wrote:

> Too often, authors of traditional review articles decide what they would like to establish as the truth either before starting the review process or after reading a few persuasive articles. Then they proceed to defend their conclusions by citing all the evidence they can find.

> The opportunity for a biased presentation is enormous, and its readers are vulnerable because they have no opportunity to examine the possibilities of biases in the review. (cited in Hunt, 1997, p. 7)

Meta-analysis, introduced in the late 1970s by the third author, was developed as an objective alternative to the traditional, narrative review (Hunt, 1997). This comprehensive statistical procedure was first applied to the question of the effectiveness of bilingual education by Okada and colleagues in 1982, but the results were not made available to the public, as mentioned previously. The first published meta-analysis of the research on bilingual education effectiveness was Ann Willig's (1985). Starting with the same corpus of studies that Baker and de Kanter (1981) had used for their narrative review, Willig sought to determine if their conclusions could be sustained using meta-analysis procedures. Willig imposed still stricter selection criteria than Baker and de Kanter, requiring that studies focus on K–12 students in U.S. schools. Willig found "positive effects for bilingual programs . . . for all major academic areas" (p. 297).

In the 1990s, Keith Baker published another narrative review, this time with Christine Rossell. 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 it is said that SI differs 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" (Chapter 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, p. 1), then it appears that the authors' SI program category overlaps in significant respects with their bilingual education program category. These imprecise definitions make it difficult to know whether a program described as "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, 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).

As Willig (1985) had done with Baker and de Kanter's (1981) narrative review, Greene (1998) conducted a meta-analysis of the studies included in Rossell and Baker's (1996) traditional synthesis, again imposing additional selection criteria, requiring that studies have measured effects after treatments lasting at least one academic year. This narrowed their 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)

In a synthesis of research on effective reading programs for ELLs, Slavin and Cheung (2003) focused on methods of teaching reading to ELL students, comparing the practice of teaching ELLs first to read in their native language (a bilingual education strategy) with that of teaching them first to read in English (an English-only 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 were used, or pretesting or other matching criteria established the degree of comparability of bilingual and immersion groups before the treatments began; (c) The subjects 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 lasted at least 1 school year. Slavin and Cheung identified 16 studies, published between 1971 and 2000, that met these criteria.

Slavin and Cheung's (2003) 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 that they found to be methodologically acceptable favored bilingual approaches over immersion approaches; although some found no difference, none favored immersion programs.

Although much work has been done to synthesize existing research on the effectiveness of bilingual education, an updated meta-analysis focusing on Arizona studies, as presented here, will provide a more focused review of evidence for state policymakers. In addition, both of these previous meta-analyses relied almost exclusively on studies conducted before 1985, while the meta-analysis reported here consists of studies conducted after 1985. Further, our perspective on meta-analysis differs from that assumed in the previous works in that we believe the broadest possible net should be cast to include as many studies as possible without applying the "best evidence" criteria of traditional narrative reviews and overly selective meta-analyses (Slavin, 1986).

Effectiveness Studies in Arizona

In this section, we discuss bilingual education effectiveness studies in Arizona[2] and two recent attempts to analyze large-scale academic achievement data in the state.

Bilingual education effectiveness research in Arizona

De la Garza and Medina (1985) studied the effects of bilingual education in a school district in Tucson. The experimental group consisted of 24 Spanish-dominant Mexican American students who had participated in a 3-year transitional bilingual education program from Grade 1 through Grade 3. The comparison group was composed of 118 English-dominant Mexican American students who had participated in a monolingual English curriculum in Grades 1–3. However, the socioeconomic status of the two groups was different. In the experimental group, 19 of the 24 students were eligible to receive free lunches, while only 41 of the 118 in the comparison group were eligible to receive free lunches. The mean scores for the bilingual-program students were higher in every grade (though the differences were not statistically significant) than for the English-dominant comparison group. Since the socioeconomic status of the comparison group is higher than that of the bilingual program group, it is likely that a controlled study would have shown still stronger results for the students who received bilingual instruction.

Two related studies (Medina, Saldate, & Mishra, 1985; Saldate, Mishra, & Medina, 1985) examined the effects of bilingual education in a school district in Douglas, a city on the Arizona–Mexican border. The two studies were conducted on a similar population, using different samples of students. Experimental and control groups in both studies were Mexican American students who spoke Spanish as a dominant language, with a matched socioeconomic status and matched scores on a language vocabulary test (Peabody Picture Vocabulary Test), who were followed from Grade 1 through Grade 3. In each study, the experimental group was enrolled in the Douglas Bilingual/Bicultural Project, and the control group was not. Achievement test scores in English and Spanish show that the comparison group scored slightly better than the bilingual education students in Grade 2, but differences were small and not statistically significant. In Grade 3, the bilingual education students outperformed the comparison students. The difference was statistically significant and quite large, and the authors suggested that the benefits of bilingual instruction became stronger over the long term.

Medina and Escamilla (1992) considered the oral English proficiency development of ELL students in two different kinds of bilingual programs in Arizona and California. Using the Language Assessment Scales in English and students' native language, Vietnamese ELL students who were enrolled in a transitional bilingual program (TBE) were assessed in California, while in

Arizona, Spanish-speaking ELLs who were enrolled in a maintenance bilingual education program (MBE), also known as developmental bilingual education (DBE), were studied. The authors report that by second grade, all students had attained comparable scores in oral English, a striking result particularly given that MBE students had been exposed to considerably less English at this point in their program. However, during this same time period, most TBE students had lost much of their native-language proficiency, while most MBE students had maintained or significantly increased their native language proficiency.

Attempts at analysis of large-scale academic achievement data in Arizona

MacSwan, Stockford, Mahoney, Thompson, and DiCerbo (2002) conducted an analysis of extant academic achievement data (Stanford Achievement Test, 9th edition [SAT-9] scores) collected by the state of Arizona; the study included the entire state population of 1,012,145 students who were in Grades 3–9 during the 5-year period from the 1996–1997 academic year through 2000–2001. As described by MacSwan and colleagues, students' scores were linked from year to year using an algorithm developed by David Garcia and colleagues, then with the Arizona Department of Education, and estimated to have 80% accuracy. The study revealed a mildly positive effect for bilingual education over students in ESL; however, the authors were not satisfied that program coding was accurate. The program placement variable, which should have remained stable, shifted erratically from one year to the next in the longitudinal data set. Due to the likelihood of errors in the data, the authors expressed concern about the reliability of the findings. The authors concluded their report as follows:

> It would appear that the LEP [limited English proficient] programs are not being carried out in the manner in which they were designed or that there may be significant miscoding of language program at the classroom or school level. The variable nature of the program information has caused us to rethink our approach to evaluating program effects in general. Assuming that the program classifications were correctly coded, we are unable to neatly characterize those experiencing TBE or ESL [English as a Second Language] programs as they are intended to be implemented. (MacSwan et al., p. 14)

The Arizona Department of Education (2004), using 2003 statewide SAT–9 scores only, conducted an analysis to determine whether students in SEI or bilingual education performed better. Immersion students in Grades 1–5, who were affected by Proposition 203, were found never to have more than a 3-month gain (measured in terms of grade equivalents) in reading and language over students in bilingual education classes. In a set of comparisons

focused on Spanish-background students, believed by the study's authors to be more accurate because the groups were less heterogeneous, students in English immersion in these grades showed either no difference or up to a 2-month advantage over students taught in bilingual education classrooms. However, the most common result for these students indicated only a 1-month advantage for English immersion over students taught in bilingual classes. Results were much more dramatic in the higher grades, where immersion students were reported to have as much as a 15-month advantage over bilingual education students in the heterogeneous group and as much as a 6-month advantage in the Spanish-background group by eighth grade.

In a critique of the Arizona Department of Education (2004) report, MacSwan (2004) noted that the study likely found progressively higher average scores for English immersion as grades progressed because it did not take students' placement histories into account, unlike the previously conducted study of statewide data (MacSwan et al., 2002), and so it had probably identified bilingual education students entering an SEI phase of their program as immersion program students; additionally, because students in higher grades in bilingual classrooms are likely to be new or recent arrivals, the study likely compared new arrivals in bilingual education classes with students with longer histories who had already transitioned to all-English classes. In addition, the data stream used in the Arizona Department of Education study was the same as that used by MacSwan and colleagues (2002), which was shown to exhibit unexpected erratic patterns of student program placements from year to year, leading to the conclusion that program placement data were often inaccurately coded or programs were incorrectly implemented. These flaws produced findings that are not meaningful.[3]

Summary of Literature Review

Research on bilingual education programs in Arizona appears to coincide with nationally situated research to support the conclusion that bilingual instructional approaches are effective for increasing ELL students' academic achievement. Nonetheless, we believe that a meta-analysis focused on Arizona studies, using effect size statistics to synthesize findings over the full range of results, will provide additional clarification by removing as much subjective judgment from the evaluation process as possible. We now turn to the present study.

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 post-1985 evaluation studies addressing programs for language-minority students. Over 300 studies were identified and reviewed.

Studies were included in the present meta-analysis according to the following selection criteria: Studies: (a) involved Arizona K–12 minority-language students (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.

Four studies meeting our selection criteria were identified: de la Garza and Medina (1985); Medina and Escamilla (1992); Medina and colleagues (1985); and Saldate and colleagues (1985). Of these, Medina and Escamilla compared one group of students in Arizona with another in California. Although one of the groups included in the study was not from Arizona, we decided to include this study as substantially meeting the criteria in the interest of enlarging our admittedly small sample. Year-to-year fluctuations in program placements suggested that a reliable description of the program models was unavailable in the extant Arizona data (MacSwan et al., 2002; Arizona Department of Education, 2004). Neither study was included in this meta-analysis due to insufficient data. MacSwan and colleagues did not provide program evaluation results due to an unreliable data set, and the Arizona Department of Education (2004) did not provide standard deviation statistics, required for the meta-analysis calculation.

Coding the Studies

Once 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 outcome measure characteristics, as shown in Table 3. 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.

Table 3

Characteristics of Studies Included in the Meta-Analysis

Study identification Author's last name Year of publication Study identification number Publication form	*Characteristics of teachers* Credentialed in bilingual education Proficient in students' language Years of experience teaching
Characteristics of program Bilingual program type Use of native language Sources of native-language support Model of native-language support Criteria used for language English proficient classification Length of time program continues in years Native-language support used for content areas	*Characteristics of research design* Type of group assignments Type of teacher assignments Control for socioeconomic status Internal validity Number of comparisons in this study
Characteristics of students Average grade level Percentage female Percentage male Socioeconomic status Ethnicity First language	*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

Calculating Effect Size

The preferred formula for estimating effect size when integrating studies that use at least two comparison groups is the difference between the mean of the first comparison group and the second comparison group on the final outcome measure, divided by the standard deviation of the second comparison group (Glass et al., 1981). All four studies were longitudinal, 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 using more bilingual education. DBE was considered to represent the most bilingual education followed, in order, by TBE, ESL/SEI, and English-only or submersion.

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 (see Table 4). We can confidently assert that the experimental group is aligned more with bilingual education pedagogy, but the program type and comparison

Table 4

Comparisons of Effect Size by Study

de la Garza and Medina, 1985			
Grades 1–3			
Range of *n*'s for transitional bilingual education (TBE): 24–25			
Range of *n*'s for English-only for non-English language learners (EO^2): 116–118			
TBE vs. EO^2	***N* of ES**	**Mean ES**	**SD of ES**
Reading vocabulary	3	0.15	0.38
Reading comprehension	3	0.17	0.06
Mathematics computation	3	-0.02	0.15
Mathematics concepts	3	-0.02	0.14
Medina and Escamilla, 1992			
Grades K–2			
Range of *n*'s for developmental bilingual education (DBE): 138			
Range of *n*'s for TBE: 123			
DBE vs. TBE	***N* of ES**	**Mean ES**	**SD of ES**
Language-oral, native	2	0.64	0.74
Language-oral, English	1	0.11	

group vary from study to study. Two studies concern the same experimental and comparison programs. Medina and colleagues (1985) and Saldate and colleagues (1985) both compare DBE to English-only for ELL students, here labeled EO[1]. But de la Garza and Medina (1985) compare TBE to English-only

Table 4, cont.,
Comparisons of Effect Size by Study

Medina, Saldate, and Mishra, 1985			
Grades 6, 8, and 12			
Range of *n*'s for DBE: 19			
Range of *n*'s for English-only for limited English proficient students (EO[1]): 24–25			
DBE vs. EO[1]	***N* of ES**	**Mean ES**	**SD of ES**
MAT test			
Total mathematics	2	-0.32	0.16
Problem solving	2	-0.24	0.13
Concepts	2	-0.34	0.25
Computation	2	-0.13	0.53
Total reading	2	-0.21	0.08
Reading	2	-0.3	0.28
Word knowledge	2	-0.1	0.1
CAT test			
Total mathematics	1	-0.2	
Concepts/application	1	-0.11	
Computation	1	-0.27	
Total reading	1	-0.63	
Comprehension	1	-0.57	
Vocabulary	1	-0.41	

Table 4, cont.,
Comparisons of Effect Size by Study

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]	***N* of ES**	**Mean ES**	**SD of ES**
Tests in English			
Total achievement	1	-0.29	
Reading	1	1.47	
Spelling	1	0.50	
Arithmetic	1	1.16	
Tests in Spanish			
Total achievement	1	0.46	
Reading	1	2.31*	
Spelling	1	3.03	
Arithmetic	1	1.16	

Note. Reading, spelling, and arithmetic are not constituents of the total achievement.

* This effect size was calculated with the treatment group's standard deviation.

for non-ELL students, here labeled EO[2]. Medina and Escamilla (1992) compare two bilingual education programs, DBE and TBE. As shown in Table 4, all outcome measures are derived from standardized tests; however, the instrument and the content area vary widely. Outcome measures used by Medina and Escamilla are different in that they are measuring oral language in both English and the native language.

When coded, the four studies yielded a total of 43 instances in which two different bilingual programs were compared on one or more outcome measures, and effect sizes were calculated for all 43. Table 4 lists the four studies and their mean effect sizes, and standard deviation for each outcome variable represented in the study. All of the studies give outcome measures, based on

standardized test scores, in English except Saldate and colleagues (1985), where it is noted that half the outcome measures are given in English and half are given in Spanish.

A positive effect size indicates that the more bilingually-instructed group did better than the less bilingual program group, while a negative effect size indicates that the less bilingually-instructed 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) compare TBE to English-only for non-ELL students who already knew English, here labeled as EO[2]. Their study shows that the size of the sample for the TBE group is about one fifth the size of the English-only group. The mean effect size for reading vocabulary is calculated as 0.15. This indicates that ELL students exposed to TBE scored one sixth of a standard deviation higher than the English-only group made up of non-ELL students. This may not seem minor as a single comparison, but taken as a whole, the accumulation of effect sizes will increase our ability to detect true differences between the various bilingual education programs and English-only environments for students.

Combining Effect Sizes for Arizona Studies

Before we report overall effect sizes for Arizona studies, some discussion is warranted concerning the integration of these results. It may be difficult to accept the integration of an oral-language test with a math test or with a reading vocabulary test, but what these outcome variables have in common is that they were selected by the researcher as a hypothesized effect of various levels of bilingual education as a program treatment. These are the effects of interest in the present meta-analysis. Even at a high level of aggregation, these effects can tell us whether bilingual programs with more native-language support are better for ELLs than those with less. Table 5 gives overall effect size results for these studies. To provide additional context, these results are compared to those of Willig's (1985) and Greene's (1998) meta-analyses of national samples of studies on bilingual education in Table 6.

Conclusions

As in previous meta-analyses conducted on national samples of effectiveness studies (Willig, 1985; Greene, 1997), our meta-analysis reveals positive effects on all measures in English, and especially positive effects for all native-language outcome measures. Moreover, of the three meta-analyses displayed in Table 6, all effect sizes are positive, and many are striking.[4] The especially high effect sizes for tests in the students' native language show the added benefits of bilingual education, which permits students to develop an ability to engage academic content in two languages.

Table 5

Combining Effect Sizes for Arizona Studies

	All outcome measures	**Reading (in English)**	**Math (in English)**	**All outcomes in native language**
Benefit of bilingual education	0.16	0.01	0.03	1.27

Table 6

Comparing the Benefit of Bilingual Education Among Meta-Analyses

	Reading (in English)	**Math (in English)**	**All outcomes in native language**
Rolstad, Mahoney, Glass (2004)	0.01	0.03	1.27
Greene (1997)	0.21	0.12	0.74
Willig (1985)	0.20	0.18	0.69

The case for bilingual education has been particularly strong in Arizona, as succinctly summarized in Table 5. It is widely conjectured that attacks on bilingual education, which putatively raise questions about its educational effectiveness, are more often ideologically driven and muddled by an array of political sentiments on issues such as immigration, social assimilation, and English-only politics (Petrovic, 1997). Hence, in Arizona and elsewhere in the United States, bilingual education appears to face opposition primarily for political, rather than pedagogical, reasons. The evaluation literature has been remarkably clear in demonstrating that bilingual education is not only as effective as English-only alternatives, but that it tends to be *more effective*.

Early proponents of English-only instruction, such as Baker and de Kanter (1981), argued for local control over questions of language of instruction, suggesting that English immersion was not an option favored at that time by the policy community. Despite this fact, English-only instruction has long been the primary, and often only, option available to children nationwide. In Arizona, which suffered from decades of English-only instruction, the limited availability of bilingual education, which became possible after the English-

only 1C program was dismantled, remained the best hope for many students, but bilingual education was never the only option. With the passage of Proposition 203 and its still more restrictive interpretation by the current superintendent of public instruction, parents and schools have now been stripped completely of their rights to choose the instructional programs that best suit children in local contexts. The irony, of course, is that the results of the empirical research indicate that Arizona has made precisely the wrong decision, banning the better of two alternatives and mandating the worse.

As shown in the present study, these policies are ill advised in light of the research evidence. In addition to the improved educational outcomes of bilingual education, such programs can be designed to treat children's native language as a resource, leading to important positive effects on self-concept, self-esteem, ethnic identification and tolerance, and development of children's native linguistic resources (Rolstad, 2000), especially in the context of two-way bilingual programs, which combine second language education for English students with maintenance bilingual education for ELLs (Rolstad, 1997). These additional and extremely important benefits are underestimated by the evaluation research, which has focused on academic outcome measures in English.

Given these research findings, it is recommended that Arizona reconsider current policies that mandate a single English-only approach for all students. Allowing students access to bilingual education programs is likely to lead to better educational outcomes for ELLs, permit the concurrent development of two-way immersion programs to serve both English-speaking and ELL students, and have additional benefits in promoting and developing students' native-language resources. Because the evaluation evidence strongly favors bilingual education over English-only alternatives, there would appear to be no rationale for banning the former—except those primarily motivated by politics and ideology.

In addition to strong research evidence, there are pragmatic reasons for supporting local, district-level decision-making about program options for English learners. SEI, as defined by Baker and others (Baker & de Kanter, 1981; Rossell & Baker, 1996), cannot be staffed statewide, as the model requires that appropriately trained teachers know the home language of students in their classrooms. Given this fact, Proposition 203 cannot feasibly be implemented, and therefore the only reasonable alternative is to permit districts flexibility in selecting program options befitting local conditions.

Furthermore, local policies could be established based not only on an estimate of what existing and newly attainable resources can support, but also on a consideration of the specific characteristics of ELL students in the local community. Hakuta and August (1998) made this point as well in the context of discussing program evaluation research on behalf of the National Research Council:

> The key issue is not finding a program that works for all children and all localities, but rather finding a set of program components that works for the children in the community of interest, given the goals, demographics, and resources of that community. (p. 147)

In Arizona, bilingual education is quite clearly an educationally effective alternative to English-only approaches, demonstrating its appropriateness to the specific communities that have used it. Policy that forbids bilingual education cannot be defended on empirical grounds.

References

Arizona Department of Education. (2000). *English acquisition services: A summary of bilingual and English as a Second Language programs for school year 98–99*. Phoenix: Author.

Arizona Department of Education. (2004). *The effects of bilingual education programs and Structured English Immersion programs on student achievement: A large-scale comparison*. Phoenix: Author.

Arizona Revised Statutes, 15–753 (2000).

Baker, C. (2001). *Foundations of bilingual education and bilingualism* (3rd ed.). Clevedon, England: Multilingual Matters.

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.

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

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), 247–249.

Dolson, D. (1985). *The applications of immersion education in the United States*. Rosslyn, VA: National Clearinghouse for Bilingual Education, InterAmerica Research Associates.

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

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237–261.

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.

Hakuta, K., & August, D. (Eds.). (1998). *Educating language-minority children*. Washington, DC: National Academy Press.

Hernández-Chavez, E. (1984). The inadequacy of English immersion education as an educational approach for language minority students in the United States. In *Studies on immersion education: A collection for U.S. educators* (pp. 144-181). Sacramento: California State Department of Education.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage Foundation.

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.

Krashen, S., Park, G. K., & Seldin, D. (2000, September/October). Bilingual education in Arizona. *NABE News*, 12–14.

Lau, J., & Chalmers, T. C. (1993). Meta-analytic stimulus for changes in clinical trials. *Statistical Methods of Medical Research, 2*, 161–172.

MacSwan, J. (2004, August 13). Bad data poison language study. *The Arizona Republic*, B9.

MacSwan, J., Stockford, S. M., Mahoney, K., Thompson, M. S., & DiCerbo, K. E. (2002). *Programs for English learners: A longitudinal exploration of program sequences and academic achievement in Arizona*. Tempe: Arizona State University and Arizona Department of Education.

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). Manuscript, Arizona State University.

Malherbe, E. C. (1978). Bilingual education in the Republic of South Africa. In B. Spolsky & R. L. Cooper (Eds.), *Case studies in bilingual education* (pp. 167–202). Rowley, MA: Newbury House.

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

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

Okada, M., Besel, R. R., Glass, G. V., Montoya-Tannatt, L., & Bachelor, P. (1982). *Synthesis of reported evaluation and research evidence on the effectiveness of bilingual education basic projects, final report: Tasks 1–6* (Contract No. 202–245–0171. Cooperative Agreement 00–CA–80–0001 for the National Institute of Education, Department of Education). Los Alamitos, CA: National Center for Bilingual Research.

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

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

Rolstad, K. (1997). Effects of two way immersion on the ethnic identification of third language students: An exploratory study. *Bilingual Research Journal, 21*(1), 43–63.

Rolstad, K. (2000). Capitalizing on diversity: Lessons from dual language immersion. *NABE News, 23*(5), 5–18.

Rolstad, K., Mahoney, K., & Glass, G. V. (In review). The big picture: A meta-analysis of program effectiveness research on English language learners.

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

Sacken, D., & Medina, M. (1990). Investigating the context of state-level policy formation: A case study on Arizona's bilingual education legislation. *Educational Evaluation and Policy Analysis, 12*(4), 389–402.

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

Secada, W. G. (1987). This is 1987, not 1980: A comment on a comment. *Review of Educational Research, 57*(3), 377–384.

Sheridan, T. (1986). *Los Tucsonenses: The Mexican community in Tucson, 1854–1941*. Tucson: University of Arizona Press.

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.

Tucker, G. R. (1980). *Implications for U.S. bilingual education: Evidence from Canadian research*. Rosslyn, VA: National Clearinghouse for Bilingual Education, InterAmerica Research Associates.

Villalobos, L. (2004, August 6). Immersion better for kids than bilingual classes, study says. *Arizona Republic*, pp. B1–B2.

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

Endnotes

[1] An extension of this project considers results from a national sample of studies (Rolstad, K., Mahoney, K., & Glass, G. V. (In review). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy*.)

[2] Our purpose is to provide a meta-analysis of research in Arizona, but we wish to refer interested readers to a narrative review of the Arizona research conducted by Krashen, Park, and Seldin (2000); these authors discussed four Arizona studies (de la Garza & Medina, 1985; Medina et al., 1985; Powers, 1978; Saldate et al., 1985) and concluded that "Arizona studies strongly suggest that bilingual education is beneficial, a conclusion that is consistent with the results of studies done in other states" (p. 5).

[3] Superintendent Tom Horne, whose political career is strongly committed to an anti-bilingual education agenda, misrepresented the results of the Arizona Department of Education's study to the press, claiming that without "a single exception," the study "tells us that the students in English immersion do substantially better" (Villalobos, 2004), a factually incorrect description of the study.

[4] To justify these descriptions of the effect sizes, we note that an effect size of .10 is approximately equal to the effect of one month's instruction in elementary school grades. That is, in reading and math tested in the elementary grades, the mean score for students in November exceeds the mean score in October by about .10 standard deviations. This is an empirical fact corroborated many times in large sets.