# What Works?

## Reviewing the Latest Evidence on Bilingual Education

*Stephen Krashen, University of Southern California*
*Grace McField, California State University, San Marcos*

The American public is under the impression that bilingual education doesn't work. Yet even a quick glance at the professional literature shows that it does. Study after study has reported that children in bilingual programs typically outperform their counterparts in all-English programs on tests of academic achievement in English. Or, at worst, they do just as well.

Numerous reviews of the research have confirmed this conclusion. This year alone, two such reviews were published (Rolstad, Mahoney, and Glass, 2005; Slavin and Cheung, 2005). And another, sponsored but not released by the U.S. Department of Education, received wide attention from the news media (see "More Evidence from the National Literacy Panel," p. 10).[1] All three found an advantage for bilingual education.

For scientists—and, one would hope, for policymakers—it is highly significant when reviews of the literature, conducted independently and examining different studies, reach similar conclusions. Such consistency provides strong evidence that research findings are reliable, rather than merely the result of chance.

It is also noteworthy that the latest reviews used a sophisticated methodology that is considered more precise and more objective than earlier approaches to summarizing research findings. The methodology is known as *meta-analysis*.

Until recently, most reviews of bilingual education research have been described as "narrative" or "vote-counting." Scholars collect a body of studies, decide which ones are worthy of inclusion, and characterize each study as favoring either bilingual or all-English programs. Then they essentially "count the votes" for each approach and declare a winner. In narrative reviews, each study—regardless of how big a difference it finds in educational outcomes, how many subjects are involved, or how rigorous its research methods—gets one vote. Up or down, yes or no, for or against bilingual education.

Several reviews of this kind have concluded that bilingual education is more effective than all-English programs in helping children to acquire English and to progress academically (Zappert and Cruz, 1977; Troike, 1978; Cummins, 1983; Krashen, 1996). On the other hand, Baker and de Kanter (1981) concluded there was no advantage (but also no harm) to bilingual education.

Alone among narrative reviews, Rossell and Baker (1996) counted more studies favoring all-English programs. Yet

they also reported only small differences between treatments and acknowledged the existence of high-quality bilingual programs.

Meta-analysis, by contrast, allows reviewers to take a more comprehensive approach. Using powerful statistical techniques, it can control for numerous variables in each study, including sample size, program model, student and teacher characteristics, research design, outcome measures, duration of study, year of publication, type of publication (e.g., dissertation, peer-reviewed journal), and so forth. These techniques can also minimize subjectivity, sometimes called "reviewer bias," in characterizing outcomes or in deciding which studies to exclude or include.

Perhaps most important, meta-analysis gives reviewers the opportunity to measure *effect size*—how big an advantage one educational treatment demonstrates over another—expressed as a single number.[2] A total effect size can be then calculated for the studies under review, as if they had all been rolled together into one big study.

Thus, meta-analysis makes it possible to reach general conclusions about the relative effectiveness of one pedagogical approach versus another. It has been suggested that an effect size of .20 represents a small impact of a treatment, while .50 represents a modest impact and .80 represents a large impact (Cohen, 1977).

**Table I:**
**Advantage for Bilingual Education in Five Meta-Analyses[3]**

| Review | Dates | Mean ES |
|---|---|---|
| Rolstad et al. (2005) | 1985 - | 0.23 |
| Slavin & Cheung (2005) | 1971 - | 0.33 |
| Willig (1985) | 1971-1980 | 0.33 |
| Greene (1997) | 1972-1991 | 0.18 |
| McField (2002) | 1968-1985 | 0.28 |
| Mean | | 0.26 |

N = number of studies covered. ES = effect size. ES of .20 = small impact; .50 = modest impact; .80 = large impact.

dence should be placed in these reviews and what overall conclusions we should draw from them.

In all studies included in these meta-analyses, students in bilingual education programs were compared with students in all-English programs. Two of the meta-analyses (Willig, 1985, and Greene, 1999) were re-analyses of vote-counting reviews (Baker and de Kanter, 1981; Rossell and Baker, 1996). Three others (Rolstad, Mahoney, and Glass, 2005; Slavin and Cheung, 2005; McField, 2002) used their own criteria in selecting a group of studies for review.

There are, of course, wide variations among bilingual programs, ranging from dual language to early-exit to concurrent translation options. There are also wide variations among programs labeled English-only, some allowing a small amount of help in the primary language, some simply "submersing" children in the mainstream, and some going to great lengths to make sure English input is comprehensible for ELLs. We do not consider such variations in this review, only comparisons between bilingual and all-English programs.

Table I summarizes five meta-analyses of studies comparing these two broad program types. Despite slightly different criteria for including studies and different dates of publication, the average effect sizes are remarkably similar, with students in bilingual education showing a small but consistently positive impact versus those in all-English classrooms.

Some caveats are in order. All of these reviews examined studies conducted in the United States only and lasting for at least one year. But one year may not be enough time for bilingual programs to show their positive effects. In most studies reviewed in the meta-analyses, comparison students and experimental (bilingual) students were ELLs. But in some studies that were included, comparison students were fluent speakers of English.

That said, the findings of the five meta-analysis were quite consistent, with a mean effect size of .26 and a range of .18 to .33. It could be argued, of course, that this was because they featured many of the same studies and are simply redundant. To determine whether this was the case, we examined studies reviewed in more than one meta-analysis. We included only those comparisons in which tests of reading comprehension in English were used, and excluded those in which fluent English speakers served as comparison students. This method not only allowed us to determine overlap, but also served as a way of measuring reliability, that is, to see whether different researchers came up with similar results.

> Study after study has reported that children in bilingual programs typically outperform their counterparts in all-English programs on tests of academic achievement in English.

## Reviewing the Reviews

We present here a "*meta*-meta-analysis," an effort to summarize the findings of published meta-analyses of programs for English language learners (ELLs). Our intent is to determine how much confi-

## Table II:
## Effect Sizes in Studies of Reading Comprehension, by Meta-Analysis[4]

| | Slavin & Cheung (2005) | Rolstad et al. (2005) | Willig (1985) | Greene (1997) | McField (2002) | Rossell & Kuder (2005) |
|---|---|---|---|---|---|---|
| Alvarez (1975) | -0.23 | | | | | - .05 |
| Huzar (1973) | 0.31 | | | 0.18 | .31, .01 | 0.16 |
| Plante (1976) | 0.5 | | | 0.52 | | 0.52 |
| Ramirez et al (1991) | | 0.01 | | 0.12 | | 0.25 |
| Campeau et al (1975) Corpus Christi | 0.45 | | | | | 0.45 |
| Maldonado (1994) | 1.66 | | | | | 0.12 |
| Campeau et al (1975) Alice | 0.49 | | | | | 0.45 |
| Saldate et al (1985) | 0.89 | 1.47 | | | 0.42 | 1.47 |
| Morgan (1971) | 0.26 | | | | 0.26 | 0.27 |
| Doebler & Mardis (1980) | 0.15 | | | | | 0.15 |
| Covey (1973) | 0.72 | | 0.74 | 0.74 | 0.74 | 0.66 |
| Medrano (1986, 1988) | | .10, -.18 | | | | |
| Kaufman (1968) | 0.23 | | 0.31 | 0.2 | .49, .11 | 0.2 |
| Danoff et al (1977) | | | 0.01 | -0.12 | | 0.12 |
| McSpadden (1979) | | | 0.2 | | | |
| Olesini (1971) | | | 0.97 | | | |
| Stebbins et. al. (1977) | | | -0.06 | | | |
| Stern (1975) | | | -0.48 | | | |
| Lindholm (1991) | | -0.59 | | | | |
| Medina, Saldate & Mishra (1985) | | -.3, -.57 | | | - .22,-13, -.51 | |
| TEA (1988) | | -0.06 | | | | |
| Powers (1978) | | | | -0.33 | -0.44 | - .35 |
| Rossell (1990) | | | | -0.05 | | -.25 |
| Bacon et al (1982) | | | | 0.68 | .82, .98 | 0.7 |
| Cohen (1975) | 0 | | | | | -.21, .08, -.28 |

Positive ES = positive effect for bilingual education; negative ES = negative effect. ES of .20 = small impact; .50 = modest impact; .80 = large impact.

Table II shows that, while there is some overlap, it is clear that all investigators did not examine the same body of research. The vast majority of studies appeared in only one or two of the five meta-analyses. So there was broad support for results favoring bilingual education.

On the other hand, when studies did appear in more than one review, there was substantial agreement about their effect size, even though effect sizes can be calculated in different ways that can produce different results. The only serious disagreement involved the effect size calculated for Saldate et al. (1985), but in all three meta-analyses the effect size was positive.

## What Kind of Bilingual Program?

In our meta-meta-analysis, we have deliberately attempted to look at the big picture to see whether there was general agreement among studies. As noted earlier, we did not attempt to examine program variations that are clearly important to pedagogical outcomes. Several individual meta-analyses, however, have attempted to do this.

Willig (1985) analyzed a number of methodological variables, reporting that studies using random assignment of subjects to experimental and comparison groups resulted in higher effect sizes favoring bilingual education. Greene (1997) reported a similar pattern. Willig also found that when comparison groups contained elements of bilingual education, such as significant use of the native language, the advantage for the bilingual program was weaker. When comparison groups contained students who had exited the bilingual program, the effect size in favor of bilingual education was considerably lower (d = -.03, versus d = .38).

## More Evidence from the National Literacy Panel

A sixth meta-analysis, funded by the U.S. Department of Education and completed in 2005, reached conclusions about bilingual education that were very similar to the five reviews described here. This study was conducted by the National Literacy Panel on Language Minority Children and Youth, a panel of researchers chosen by the Bush Administration.

Last summer, however, the Administration decided against releasing the panel's report, even before it was completed. Grover Whitehurst, director of the federal Institute for Education Sciences, complained that the study had methodological and editorial problems.

"What we got," Whitehurst told Education Week, "was a report that would be a useful work on the bookshelf of researchers who spend all their time on this topic, but it was too long and inaccessible to be useful to practitioners." The Department of Education, which had spent $1.8 million on the study, chose to abandon it while the panel was still editing its final report.

The decision raised questions about whether it was really the panel's conclusions—favoring bilingual education over all-English approaches in teaching ELLs to read—that worried the Bush Administration. While the White House has not actively opposed native-language instruction, many of its supporters are ideologically committed to English-only policies.

The National Literacy Panel study reportedly found a small to modest impact for bilingual versus nonbilingual programs. And, like other meta-analyses, it noted that the most rigorous research designs—those that used random assignment—showed the biggest edge for bilingual education.

Rather than release these findings officially, the Department of Education has agreed to surrender the copyright and allow the panel to publish its study privately. The results will then be available to other researchers. Their impact on policymakers and practitioners, however, is likely to be far less than it would have been if the Department had endorsed the report.

*– James Crawford*

Willig concluded that positive effects for bilingual education were apparent only when methodological weaknesses in the studies were controlled. In other words, the better the research design, the stronger the effects for bilingual education.

Others have investigated the impact of the kind of bilingual program used. McField (2002) concluded that programs designed along principles hypothesized to underlie ideal bilingual programs (e.g., Krashen, 1996) were more effective. But very few such comparisons were possible (only one "strong" program and four "weak" programs could be analyzed in this way). Rolstad, Mahoney, and Glass (2005) present evidence suggesting that late-exit, developmental bilingual programs are more effective than early-exit, transitional programs.

According to Cohen's (1977) standard, the average effect size for bilingual education is small. But the strikingly similar results from different meta-analyses provide clear support for bilingual education as a means of helping children succeed academically in English. They also cast strong doubt on claims that all-English approaches are superior and should be mandated by law.

There is no doubt that, when it comes to English acquisition, native-language instruction is part of the solution, not part of the problem. As research continues to yield information about the factors that predict successful programs for ELLs, it is likely that we will see larger effect sizes for bilingual education in the future.

## References

*Bibliographic references are not included for the studies in Table II. They are available in the meta-analyses listed below.*


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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cummins, J. (1983). *Heritage language education: A literature review.* Toronto: Ministry of Education.

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

Greene, J. (1997). A meta-analysis of the Rossell & Baker review of bilingual education research. *Bilingual Research Journal* 21(2/3): 103-122.

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

McField, G. (2002). Does program quality matter? A meta-analysis of select bilingual education studies. Ph.D. Dissertation, University of Southern California.

Rolstad, K., Mahoney, K., and Glass, G. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy* 19(4): 572-594.

Rosenthal, R. (1984). Meta-analysis for social research. Beverly Hills, CA: Sage.

Rossel, C. (1990) The effectiveness of educational alternatives for limited English-proficient children. In G. Imhoff (ed.), *Learning in two languages* (pp. 71-121). New Brunswick: Transcation.

Rossell, C., and Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English 30*(1): 7-69.

Rossell, C., and Kudar, J. (2005). Meta-murky: A Rebuttal to recent meta-analyses of bilingual education. In J. Sohn (Ed.), *The effectiveness of bilingual school programs for immigrant children* (pp. 43-76). Berlin: AKI.

Slavin, R., and Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research* 75(2): 247-284.

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

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


*Notes available on page 34* ▶▶▶

## Learn More at NABE 2006

**Authors of two recent meta-analyses will discuss their work at the upcoming NABE conference in Phoenix, January 18-21, 2006:**

**Kellie Rolstad**, **Kate Mahoney**, and **Gene Glass** of Arizona State University will present on their research conducted at Arizona State University.

**Diane August** of the Center for Applied Linguistics and David Francis of the University of Houston will provide details of the National Literacy Panel study.

# What Works? Reviewing the Latest Eveidence from Bilingual Education

▶▶▶ *continued from page 10*

## Notes

1. Because the federal study has yet to be published, it is not among those reviewed in this article.

2. For the statistically minded, the usual way of computing effect sizes is to subtract the mean posttest score of the comparison group (in this case, all-English) from the mean posttest score of the experimental group (in this case, bilingual education), and divide by the pooled standard deviation. For example, if students in a bilingual program scored 80 on a reading test after three years of bilingual education, students with similar backgrounds in an all-English program scored 70, and the average standard deviation on the test was 10, then the effect size would be 1.0 (80 minus 70, divided by 10). Of course, computing effect sizes is not always that simple. Pretest scores sometimes differ, crucial information is sometimes lacking from published reports, and there are a variety of other ways of computing effect sizes (Glass, McGaw, and Smith, 1981; Rosenthal, 1986).

3. The effect sizes are for all measures of achievement combined, except for Slavin and Cheung (2005), who considered only tests of English reading. Most reviewers included only studies in which students were randomly assigned to treatments or in which other means of matching students were used. Rolstad, Mahoney, and Glass (2005) did not feature this requirement.

   Rossell and Kudar (2005) arrived at an average effect size of .14 for the studies covered in Slavin and Cheung, limiting their analysis to studies of Spanish-speaking children in elementary school (14 studies). They also calculated an average effect size of -.07 for Greene's studies using reading as a measure, compared to Greene's result of .21 for reading. Effect size calculations for most individual studies were very similar, but Rossell and Kudar calculated an effect size of -.25 for Rossell (1990), claiming that Greene did not use the final year of the study. We estimated an effect of size .10 for that year, based on Rossell's regression results (from Rossell, 1990, appendix 2). Using a sample expanded by adding chance scores for students eligible for the test but who did not take it, the effect size moves to a negative 1.66 (data in Rossell, p. 91, Table 4.6).

4. McField (2005) considered separate cohorts, hence the presence of more than one effect size in some cases. Gersten's studies (from Rolstad, Mahoney, and Glass, 2005) are not included; for discussion, see Krashen (1996). Rossell and Kuder (2005) note that Gersten (1985) did not involve bilingual education. In Lindholm (1991), the effect size was based only on grade 2; there was no significant difference between bilingual and comparison students in grade 3 but it was impossible to compute effect sizes from the information provided. The Medrano (1986) effect size is based on grade 6 results. See Medrano (1988) for grade 3 results.

   Rossell and Kuder consider Maldonado (1994) to be an "outlier" because the effect size is "unbelievable." They note that the exceptionally large effect size could have been due at least in part to teacher differences: "[T]he teacher assigned to the treatment group had experience working with 'integrated bilingual special education' and teaching bilingual students with learning disabilities. The control group teacher apparently had no experience working with bilingual students with learning disabilities … The teaching strategies used by the experimental group teacher [also] include a wide range of strategies beyond the language of instruction" (p. 56). In addition, the gains made by the experimental group were so "astonishing" that Rossell and Kuder say that "one can only wonder if the researcher made a mathematical or other kind of error" (p. 59).