A Longitudinal Model for Interpreting Thirty Years of Bilingual Education Research

Jesús José Salazar
University of Southern California Title VII Fellow

Abstract

This article summarizes the results from the re-analysis of two recent bilingual education studies: Greene's (1998) meta-analysis and Thomas and Collier's (1997) longitudinal study. First , the article presents the ten main criteria identified over the past three decades by the research community for conducting a methodologically adequate bilingual education study. Second, based on a power analysis of Greene's data, it is argued that a major reason why many bilingual education studies over the past thirty years remain uninterpretable is because of the high occurrence of Type II errors. Third, based on an effect size analysis of Thomas and Collier's data, it is maintained that their longitudinal model best explains thirty years of bilingual education research. When Thomas and Collier's data are converted into effect sizes, their model becomes a longitudinal meta-analysis, and their findings support the results from Greene's and Willig's (1985) meta-analyses. This article concludes by making recommendations for enhancing the evaluation of Title VII programs.

The effectiveness of Title VII programs needs to be examined in the context of 30 years of bilingual education research. The purpose of this article is to report the results of the re-analysis of two well known bilingual education studies with the goal of assisting educators to interpret Title VII program outcomes. First, the article presents the results of a power analysis of the studies included in Greene's (1998) meta-analysis of bilingual education programs. This power analysis reveals that most bilingual education reading studies in the past 30 years have been so statistically weak that the results have generally been uninterpretable. Second, the article presents the outcomes of a re-analysis of Thomas and Collier's (1997). 15-year longitudinal study. It is argued that the Thomas and Collier model best explains three decades of bilingual education research and also serves as a heuristic framework for measuring and improving the effectiveness of Title VII programs.

Before discussing the results of the re-analysis of Greene's and Thomas and Collier' data, an overview of the methodological criteria for conducting appropriate bilingual education research is provided. The summary of these research design issues sets the framework for discussing the specific studies identified above.

Research Focus on Short Term Effects and Design Limitations

After three decades of bilingual education research, we are no closer to addressing the long-term effectiveness of bilingual education programs for English language learners (ELL) as compared to English-only programs. A major reason for the inconclusive results is that most bilingual education studies conducted to date have been *short-term*. According to Greene (1998), the average bilingual education study covers a two-year period. Willig (1984) noted that about two-thirds of the studies she reviewed covered a period of one year or less. Short-term studies may subsequently lead to short-sighted conclusions about the effectiveness of both primary language and English-only programs. Moreover, Thomas and Collier (1997) are not scheduled to release their complete longitudinal outcomes for another three years. Therefore, the *long-term* educational effects of bilingual education programs, compared with English-only programs, are not truly known yet.

The on-going debate in the bilingual education research literature ultimately centers on two issues. The primary issue is: under what conditions do bilingual education programs result in greater English achievement gains for English learners than English-only programs, and vice versa. However, before this critical question can be fully answered, a second question also needs to be answered: what constitutes a good bilingual education research study? This is not a trivial question because of the more than one thousand bilingual education studies conducted to date, nearly 90 percent (Lam 1992, Salazar 1998) have been rejected as methodologically inadequate by the research community.

Rossell and Baker's (1996) review and Greene's (1998) meta-analysis provide the clearest criteria for accepting or rejecting studies as methodologically adequate. Rossell and Baker identify four methodological criteria for judging the adequacy of a study (first four criteria listed below). Greene concurs and adds a fifth guideline for identifying valid and reliable studies. Crawford (1999) and Krashen (1999) identify the sixth guideline for adequate bilingual education studies. I add two additional criteria (the seventh and eighth criteria listed below). The eight criteria are:

1. Compare students in a bilingual program to a control group of similar students.
2. Differences between the treatment and control groups have to be controlled statistically or randomly assigned to treatment and control groups.
3. Results have to be based on standardized test scores in English.
4. Differences between the scores of treatment and control groups have to be determined by applying appropriate statistical tests.

5. Only studies that measure the effects of bilingual programs after *at least one academic year* should be included in any review of the literature.
6. Exclude studies of Canadian immersion programs since they do not compare the *acquisition of English* by French Canadians with the *acquisition of English* by English language learners.
7. Exact length of treatment in the primary language program has to be documented and reported.
8. Effect size difference between the bilingual education and English-only programs should be calculated and reported.

Thomas and Collier (1997) recommend two additional criteria for evaluating the adequacy of bilingual education research. First, they suggest that the only good bilingual education study is a longitudinal study. Second, the only good longitudinal study is one that follows a cohort's long-term academic gains beyond the elementary school grades and into the middle and high school grades.

Re-Analysis of Greene's (1998) Data and Type II Errors

Salazar (1998) performed a power analysis of Greene's (1998) meta-analysis data to highlight the existence of an exorbitant rate of Type II errors in the bilingual education research literature. As Salazar notes, the chances of making a Type II error in bilingual education research is about 58 percent (see Table 1) in studies of English reading (based on the ten English reading studies included in Greene's (1998) meta-analysis). This is nearly three times greater than the accepted rate of 20 percent for Type II errors (Cohen, 1988). In other words, there is only about a one in three chance of detecting statistically significant differences in English reading between bilingual education and English-only programs. One has a better chance of obtaining a statistically significant difference by simply flipping a coin, where the odds improve to 50 percent.

A power analysis is a statistical method for calculating the probability of detecting a statistically significant outcome. As defined by Cohen (1988), the accepted power should be no less than .80. In other words, the probability of making a Type II Error should be no greater than .20 (see also Welkowitz, Ewen & Cohen, 1991). As Table 1 illustrates, there is only a 42% chance of detecting a significant difference in reading between bilingual education and English-only programs, which is only half the 80 percent statistical power recommended by Cohen.

The results of this power analysis correspond to the findings of "no differences" reported in the research literature. As noted in Greene's meta-analysis, the odds are 58 percent (Type II Error rate in Greene's data) that we will find no significant differences between bilingual education and English-only programs. Indeed, Baker and de Kanter (1981), Rossell and

Table 1. Power Analysis of English Reading Studies Included in Greene's (1998) Meta-Analysis

Study	Effect Size	Sample Size	Statistical Power	Type II Error Rate (%)
Bacon (1982)	.68	36	.15	85
Covey (1973)	.74	36	.50	50
Danoff (1977)	-.12	1488	.99	1
Huzar (1973)	.18	86	.28	72
Kaufman (1968)	.20	74	.27	73
Plante (1976)	.52	28	.11	89
Powers (1978)	-.33	87	.28	72
Ramirez (1991)	.12	248	.65	35
Rossell (1990)	-.05	347	.79	21
Skoczylas (1992)	.13	50	.18	82
Average	.21	262	.42	58

The effect sizes were calculated by Greene (1998) in his meta-analysis study. The statistical power and the Type II error rate were calculated by utilizing Cohen's (1998) statistical power analysis techniques which employ each study's sample size and Greene's average efffect size of .21 standard deviations.

Baker (1996), and Rossell and Ross (1986) report that between 45 percent and 50 percent of the studies comparing English reading outcomes between bilingual education and English-only programs show no significant difference. The number of studies reporting "no differences" in math between the two programs is even higher, ranging between 56 percent and 67 percent. These findings of "no differences" may thus be more a function of low statistical power (high Type II error rate) than actual lack of differences between the bilingual education and English-only programs.

The significance of the power analysis of Greene's data is that it points to one of the main reasons why we are no closer to addressing the long-term effectiveness of bilingual education programs compared with English-only programs. That is, most studies have used statistically weak designs that have failed to detect program differences. The odds (58 percent Type II Error rate) have been stacked against finding significant differences. The lesson to be learned by current and future Title VII

projects is that the evaluators of those projects need to estimate the power of their respective studies and the sample size that will be needed to increase each study's power, thereby reducing the risk of Type II Errors.

The Thomas and Collier Study: A Reanalysis

There has been much confusion on how to interpret Thomas and Collier's (1997) longitudinal outcomes for two reasons. First, it is not clear if their data represent actual K-12 longitudinal achievement scores, as they claim in the body of their text, or if the data represent "Results aggregated from a series of 4-8 year longitudinal studies (p. 53)" as stated in the caption of their famous chart. Second, Thomas and Collier have been faulted for not presenting complete tables of samples and descriptive and inferential statistical results. Their minimalist approach to data reporting may be best described as "research lite." Their work has been criticized as long on conclusions but short on the facts. One of the biggest critics of bilingual education has even questioned the veracity of their data (Rossell 1998).

However, Thomas and Collier provide enough information to convert their original results into effect size differences between primary language programs and English-only programs. They provide the NCE (normal curve equivalent) scores for English reading and the corresponding standard deviation of 21.06. Based on these NCE scores and the standard deviation, their data have been converted into effect sizes, as depicted in Table 2. Once the Thomas and Collier data are converted into effect sizes, they can be interpreted as a meta-analysis study. In fact, this re-analysis of Thomas and Collier's data may best be described as a *longitudinal meta-analysis*. Their results can therefore be compared with the Willig (1985) and the Greene (1998) meta-analyses, as described below.

An effect size is the standard deviation difference between a treatment group (bilingual program) and a comparison group (English-only program). In short, an effect size is a standardized score. The effect sizes computed in this re-analysis of Thomas and Collier's longitudinal data were calculated by subtracting each primary language program's (treatment group) NCE score from the English-only (comparison group) NCE score, then dividing the difference by the standard deviation of 21.06. The English-only comparison group in the Thomas and Collier study comprised English language learners who received their instruction through either a content-based ESL or English structured immersion program (English-only programs). It needs to be emphasized that, aside from the mean NCE scores and standard deviation, Thomas and Collier have presented no other data, including the sample sizes studied.

Table 2 shows the unique and differential effect sizes of each primary language program in relation to the English-only comparison group *across all grade levels*. Table 3 shows the effect size differences of various combinations of primary language programs in relation to the comparison

group. This re-analysis of Thomas and Collier's longitudinal data strongly suggests that one's *research results may be strongly influenced by the grade at which the data are collected.*

Table 2 shows each primary language program from the most effective to the least effective in the acquisition of English reading skills, compared with the English-only comparison program. However, in previous reviews of the bilingual education research literature, primary language programs have been lumped together, thereby masking the unique effects of the more successful programs with the effects of the least successful programs. Table 3 thus depicts different combinations of primary language programs and their differential effects in the acquisition of English reading skills. That is, Table 3 depicts the combined effects of various groups of primary language programs, from the most effective combination of programs to the least effective program combinations. Table 3 therefore more closely represents the state of affairs in the literature when the effects of the various primary language programs are grouped together, such as in meta-analysis studies (Greene 1998, Willig 1985) or vote-counting studies (Baker & de Kanter 1981, Rossell & Baker 1996, and Rossell & Ross 1986).

As noted, the data displayed in Table 2 and Table 3 depicting effect size differences between primary language programs and English-only programs may also be interpreted as a meta-analysis of Thomas and Collier's longitudinal data. As such, we can compare the Thomas and Collier data with the Willig meta-analysis and the Greene meta-analysis outcomes. Willig reported an average effect size gain in English reading of .20 standard deviations, in favor of primary language programs when compared with English-only programs. Greene reported a nearly identical average effect size gain of .21 standard deviations in favor of primary language programs. The meta-analysis results of Thomas and Collier's data from the last row of Table 2 (All Programs Combined) shows an overall effect size gain of .26 standard deviations in favor of primary language programs. This overall effect size result noted in Thomas and Collier's data is therefore comparable to the meta-analysis outcomes reported by Willig and by Greene.

In concrete terms, a difference of one-fifth to one-fourth of a standard deviation means that if a group of English language learners in a primary language program and a group in an English-only program both began the school year at the 30th percentile—and if those students in the English-only program maintained their English reading score at the 30th percentile the following school year—those students in the primary language program would now be at the 38th percentile. An effect size difference of this magnitude, therefore, translates into educationally significant gains.

The Thomas and Collier Model: An Evaluative Framework

As Table 2 illustrates, the effect size differences between the respective primary language programs and the English-only comparison group vary by *program type* and by *grade*. The re-analysis of the Thomas and Collier data

Table 2. Unique Effect Size Differences Between Primary Language Programs and English-Only Programs in English Reading*

Program	Grade											
	1	2	3	4	5	6	7	8	9	10	11	Ave
Two-Way	0	0	-.09	.24	.24	.47	.66	.85	1.04	1.18	1.28	.53
Late Exit	0	0	-.14	.29	.29	.28	.42	.66	.71	.80	.85	.38
Early Exit	0	0	0	0	0	.05	.05	.14	.19	.23	.28	.09
Concurrent	0	0	0	0	.05	.05	.05	.05	.05	.05	.05	.03
Programs Combined	0	0	-.06	.13	.15	.21	.30	.43	.50	.57	.62	.26

**The effect sizes were calculated by subtracting each bilingual program's NCE (normal curve equivalent) score from the comparison group NCE, and dividing by the national norm standard deviation of 21.06. Thomas and Collier prefer to use the national norm standard deviation rather than local or state standard deviations because they are more conservative (local district standard deviations have less variability, and consequently can produce inflated effect sizes).*

clearly shows that research results will be strongly influenced by both the type of program and the grade at which the data are collected. Herein lies the heuristic value of utilizing Table 2 and Table 3 to interpret both current and previous research of bilingual education programs. Thomas and Collier's longitudinal findings, when translated into effect sizes, provide an evaluative framework for explaining bilingual education research outcomes for the past three decades. The Thomas and Collier model can also be used to predict the likely outcomes of current bilingual education programs, including Title VII programs being funded by the U.S. Department of Education through the Office of Bilingual Education and Minority Languages Affairs (OBEMLA).

As Tables 2 and 3 illustrate, there are little or no effect size differences between primary language programs and English-only programs in Grades K-3. In fact, Thomas and Collier's data indicate that results through Grade 3 actually favored the English-only comparison group when compared with two-way bilingual programs and late-exit bilingual programs. Thomas and Collier's model therefore predicts that *one should expect negative outcomes from certain types of primary language programs in the early elementary grades* when compared with English-only programs. The effectiveness of late-exit and two-way bilingual programs become evident in the middle school years and continue to widen throughout the high school grades when compared with English-only programs. However, since the average bilingual education study covers a two-year period, these longitudinal effects are completely overlooked under the current short-term research paradigm.

Table 3. Combined Effect Size Differences between Primary Language Programs and English-Only Programs in English Reading*

Program	Grades											
	1	2	3	4	5	6	7	8	9	10	11	Ave
Two-Way & Late Exit	0	0	-.11	.27	.27	.38	.53	.80	.88	.99	1.06	.37
Two-Way, Late Exit & Early Exit[1]	0	0	-.08	.18	.18	.27	.38	.55	.65	.74	.80	.33
Two-Way & early Exit[1]	0	0	-.04	.12	.12	.26	.35	.49	.61	.71	.78	.31
Two Way & Early Exit[2]	0	0	-.04	.12	.14	.26	.35	.45	.54	.61	.66	.28
Late-Exit & Early Exit[1]	0	0	-.07	.14	.14	.16	.23	.40	.45	.51	.56	.23
Two-Way, Early-Exit[1] & Early Exit[2]	0	0	-.03	.08	.10	.19	.25	.35	.43	.49	.54	.22
Late-Exit & Early-Exit[2]	0	0	-.07	.14	.17	.16	.23	.35	.38	.42	.45	.20
Late Exit, Early Exit[1], & Early Exit[2]	0	0	-.04	.10	.11	.13	.17	.28	.32	.36	.39	.17
Early Exit[1] & Early Exit[2]	0	0	0	0	.02	.05	.05	.09	.12	.14	.16	.06
All Programs Combined	0	0	-.06	.13	.15	.21	.30	.43	.50	.57	.62	.26

[1]Early-exit bilingual programs with either SDAIE (Specially Designed Academic Instruction in English) or ESL taught through academic content.
[2]Early-exit bilingual programs with ESL taught traditionally.

**The effect sizes were calculated by subtracting each bilingual program's NCE (normal curve equivalent) score from the comparison group NCE and dividing by the national norm standard deviation of 21.06. Thomas and Collier prefer to use the national norm standard deviation rather than local or state standard deviations because they are more conservative (local district standard deviations have less variability, and consequently can produce inflated effect sizes).*

That research results may be greatly influenced by the program type and by the grade at which the data are collected has enormous implications for how we interpret the bilingual education research literature. Since the majority of studies comparing the effects of bilingual education programs with those of English-only programs have been conducted in the elementary school grades, these studies will, therefore, generally fail to capture any longitudinal gains that may occur in the middle school and high school grades.

In the remainder of this article, the Thomas and Collier model is utilized to further explain the outcome implications found in the two most common bilingual education programs: early-exit and late-exit programs.

Early-Exit Bilingual Programs

Early-exit bilingual programs comprise the most often implemented primary language program for English language learners, as noted by Thomas and Collier (1997). Early-exit bilingual programs are by definition *early elementary school* programs, with English language learners exiting the primary language program by no later than the third grade (Thomas & Collier, 1997). According to the Thomas and Collier model, one is guaranteed to find no differences in English reading achievement between early-exit bilingual programs and English-only programs in the elementary school grades. According to their model, neither is one likely to find any English reading differences in the middle school grades. Their model predicts that differences in English reading do not become evident until either the eighth grade or the beginning of high school.

In short, early-exit bilingual programs will inevitably "prove" to be inconclusive from Kindergarten through about the eighth grade. This is why one of the more celebrated studies by Ramirez, Pasta, Yuen, Billings, and Ramey (1991) found no differences in English achievement between bilingual education and English-only programs. Since Ramirez et al. followed the bilingual education and English-only programs only through the third grade, the Thomas and Collier model would have predicted no program differences. Unless one is committed to following early-exit bilingual education program and English-only program cohorts longitudinally through the high school grades, one will not likely find significant differences between the two programs any time soon. The positive effects of early-exit bilingual programs will not become evident until the high school grades. Since none of the early-exit bilingual programs reported in the research literature have tracked student progress from the elementary grades and into the high school grades (with the exception of Thomas and Collier's study), the longitudinal effects reported by Thomas and Collier have not been replicated.

Late-Exit Bilingual Programs

In late-exit bilingual programs, English language learners transition into English-only classrooms usually by the fifth grade. The transition into English-only instruction is thus more gradual than in early-exit programs. The rationale for late transition into English-only classrooms is based on Cummins' (1981) claim that it takes between five and seven years for English language learners to become proficient in English academic skills. The transition from primary language instruction to English-only instruction is therefore completed by the late elementary grades.

While the Thomas and Collier model predicts "no difference" in English reading between primary language and English-only instruction programs in the elementary school grades, their model predicts initial "negative" outcomes for bilingual education programs at the elementary level. It indicates that English reading outcomes through Grade 3 actually favor English-only programs compared with late-exit bilingual programs (effect size difference of .14 standard deviations in favor of English-only program).

Critics of bilingual education (Baker & de Kanter 1981, 1983; Rossell & Baker, 1996; Rossell & Ross, 1986) repeatedly mention that a greater number of studies show that English-only programs are more effective than bilingual education programs in producing greater English academic gains. These critics do not always cite the specific studies and the corresponding bilingual programs being evaluated. Thomas and Collier model predicts that the bilingual programs that are identified as less effective than English-only programs are most likely late-exit programs during the early grades (K-3). The critical reviews of bilingual education by Baker and de Kanter (1981, 1983), Rossell and Baker (1996), and Rossell & Ross (1986) need to be replicated to verify if the least effective elementary school bilingual programs were indeed late-exit programs.

Thomas and Collier's model also indicates that the positive effects of two-way and late-exit bilingual programs do not become evident until the upper elementary grades (Grades 4 or 5). The effect size differences favoring these two primary language programs dramatically increase in the middle school years, and continue to widen throughout the high school grades.

Recommendations

Based on the re-analysis of the Thomas and Collier longitudinal study, it is strongly recommended that the Office of Bilingual Education and Minority Languages Affairs adopt a longitudinal framework in the evaluation of Title VII programs. For example, the Title VII Comprehensive Grants that are funded for a five-year period should also require a five-year longitudinal evaluation of each funded program. Selected Title VII Comprehensive Grants can be refunded for an additional two years

through the Title VII Enhancement Grants, thereby making it possible to evaluate the seven-year longitudinal effects of primary language programs, compared with English-only programs.

OBEMLA needs to develop an evaluation framework that all grantees should follow. In other words, a uniform evaluation design should be constructed that will assure the validity and reliability of a program's evaluation, while also ensuring that each program is uniquely evaluated. The OBEMLA evaluation guidelines should include a required power analysis to minimize Type II errors. Moreover, the guidelines should require calculation of a program's effect size in order to measure each Title VII project's effectiveness against effect size benchmarks already established by Greene, Thomas and Collier, and Willig. To date, most bilingual education research has been uninterpretable due to the lack of a longitudinal framework and to the high occurrence of Type II errors. OBEMLA stands at the crossroads where it can take corrective action to ensure quality evaluations of primary language programs by minimizing Type II errors and encouraging a long-term focus in the evaluation of both primary language and English-only programs.

References

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

Crawford, J. (1999). Life in a political climate: What role for educational researchers. Paper presented at the Linguistic Minority Research Institute at the Conference on the Schooling of English Language Learners in the Post 227 Era, Sacramento, CA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cummins, J. (1981). Primary language instruction and the education of language minority students. *In Schooling and language students: A theoretical framework,* Los Angeles, CA: Evaluation, Dissemination and Assessment Center, California State University.

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

Krashen, S. (1999). *Condemned without a trail: Bogus arguments against bilingual education.* Portsmouth, NH: Heinemann.

Lam, T. C. M. (1992). Review of Practices and Problems in the Evaluation of Bilingual Education. *Review of Educational Research*, 2, 181-203.

Ramirez, J.; Pasta, D.; Yuen, S.; Billings, D. & Ramey, D. (1991). *Final report: Longitudinal study of structured English immersion strategy, early-exit and late-exit transitional bilingual educational programs for language minority children*, Vol. I-BII. San Mateo, CA: Aguirre International.

Rossell, C. H. (1998). *Mystery on the bilingual express: A critique of the Thomas and Collier Study*. Boston, MA: The READ Institute, Boston University.

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

Rossell, C. H. and J. M. Ross (1986). The social science evidence on bilingual education. *Journal of Law and Education*, 15, 385-418.

Salazar, J. J. (1998). *Selected Review of the Bilingual Education Research Literature*. Los Angeles: USC Graduate School of Education, Unpublished Manuscript.

Thomas, W. P. & V. Collier (1997). *School Effectiveness for Lanuage Minority Students*. Washington, DC: National Clearinghouse for Bilingual Education.

Welkowitz, J., R. B. Ewen, and J. Cohen (1991). *Introductory statistics for the behavioral sciences (4th Edition)*. San Diego, CA: Harcourt Brace Jovanovich.

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